JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Volume 12, No. 2
August 2019
ISSN: 1946-1836

In this issue:

4. **A Machine Learning Approach to Optimizing Diabetes Healthcare Management Using SAS Analytic Suite**
Taiwo Ajani, Ferrum College
George S. Habek, North Carolina State University

18. **Aligning the Factors of Skill, Technology and Task to Influence the Continued Use of Mobile Payment Users in China**
Charlie Chen, Appalachian State University
Hoon Seok Choi, Appalachian State University
Jason (Jie) Xiong, Appalachian State University
Yong Tang, University of Electronic Science and Technology

30. **Security, Privacy, and Legislation Issues related to Commercial Drone Deliveries**
Sandra A. Vannoy, Appalachian State University
Dawn Medlin, Appalachian State University

37. **Effects of Normalization Techniques on Logistic Regression in Data Science**
Adekunle Adeyemo
Hayden Wimmer, Georgia Southern University
Loreen Powell, Bloomsburg University

iscap
Information Systems & Computing Academic Professionals

The **Journal of Information Systems Applied Research** (JISAR) is a double-blind peer reviewed academic journal published by ISCAP, Information Systems and Computing Academic Professionals. Publishing frequency is three issues a year. The first date of publication was December 1, 2008.

JISAR is published online (http://jisar.org) in connection with CONISAR, the Conference on Information Systems Applied Research, which is also double-blind peer reviewed. Our sister publication, the Proceedings of CONISAR, features all papers, panels, workshops, and presentations from the conference. (http://conisar.org)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%), other journal papers (top 30%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the JISAR journal. Currently the target acceptance rate for the journal is about 40%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org. Special thanks to members of AITP-EDSIG who perform the editorial and review processes for JISAR.

JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Editors

2019 JISAR Editorial Board

A Machine Learning Approach to Optimizing Diabetes Healthcare Management Using SAS Analytic Suite

Taiwo Ajani
tajani@ferrum.edu
Computer Technology and Information Systems
Ferrum College
Ferrum, VA 24088

George S. Habek
ghabek@ncsu.edu
Business Analytics
North Carolina State University
Raleigh, NC 27695

Abstract

A typical healthcare business challenge was explored to demonstrate the effectiveness of data mining and machine learning techniques on large-scale medical and pharmacy claims data for about 70,000 patients newly diagnosed with type II diabetes. The business challenge was to move uncontrolled diabetic patient (H1AC > 7) to a controlled state (H1AC < 7). Two models were explored for this purpose and the regression was observed to perform slightly better than decision tree. Regression model was subsequently used to score "new" data. Analyses revealed the drivers and probabilities of a patient being diagnosed as controlled. Obtained results can provide incentives for the business decision maker to explore interventional programs that could enhance the quality of treatment for the uncontrolled diabetic. The article provides an added value to business and the analytic literature by exploring and explaining predictive analytics and associated techniques from the perspective of the business.

Keywords: data mining, healthcare analytics, SAS suite, machine learning, regression, decision tree.

1. INTRODUCTION

The healthcare landscape is experiencing unprecedented growth in data complexity influenced by rapid technology changes, availability of mobile applications, newly discovered diseases and evolving legislation, However, rapid progress is being made in clinical and data analytics, with unprecedented opportunities for improving healthcare quality (Bates et al., 2014). Healthcare systems are interested in predicting who will become very sick, return to the emergency room (ER) or even die after a recent diagnosis or treatment. As a result, the sector is increasingly interested in exploring new techniques to manage complex data for pattern discovery and decision making (Alkhatib, Talaei-Khoel & Ghapanchi, 2015; Raghupathi, 2010).

Data mining has been proposed in several quarters to address existing data deluge with the proposition that anecdotal data can be trained, validated, modeled and used to score "new" data (Saini & Kohli, 2018; Alkhatib, Talaei-Khoel & Ghapanchi, 2015; Raghupathi & Raghupathi, 2014; Koh & Tan, 2005). Alkhatib (2015, p.3) described data mining as a "process by which data is gathered, analyzed and stored in order to

produce useful and quality information and knowledge". According to Srinivas et al. (2010), mining is an important step in a knowledge discovery process that consists of data cleaning, data integration, data selection, pattern recognition and knowledge presentation. In contrast to traditional data analysis, data mining is discovery driven and may accomplish class description, association, classification, clustering, prediction and time series analysis (Srinivas, Rani, & Govrdhan, 2010).

The purpose of data mining and analytics is to build models for determining an event and predicting the outcome of it. This allows decision makers an actionable information upon which interventional approaches can be developed and deployed. Several authors have described classification data mining techniques with illustrations of their applications to healthcare; these include: *Rule set classifiers*, *Neuro-Fuzzy* (Srinivas et al., 2010), *Bayesian Network Structure Discoveries, Decision Tree algorithms*, *Neural Network Architecture, Support Vector Machines* (Hachesu, et al. 2013). Others have used *Regression algorithms* in healthcare analytics. This case study uses SAS suite to select the better of decision tree (DT) and regression (R) models in analyzing drivers of a controlled diabetic.

Models: Decision Tree and Regression

According to Suryawanshi, DT uses a combination of mathematical and computational techniques to aid description and classification (2012), and to extract knowledge from datasets (Kaur & Wasan, 2004). It is a visual and analytical decision support tool that can provide a graphical representation of obtained knowledge in the form of a tree in a flow chart-like structure (Kaur et al., 2018; Kajabadi, 2009; Kaur & Wasan, 2004). Kajabadi et al. (2009) wrote that each non-leaf node denotes a test on an attribute, and each branch indicates an output of the test. As a result of the hierarchical arrangement of nodes and branches, DT is easily understandable and interpreted. Kaur and Wasan (2006) posited that DTs are best suited for data mining because they are inexpensive to construct, easy to interpret, easy to integrate with database systems and have comparable or better accuracy in many applications. Its reliability and established accuracy in clinical decision-making make it the technique of choice for this study (Sitar-Taut & Sitar-Taut, 2010).

Although the DT can be applied as a *continuous* analytic instrument, it is naturally suitable to address discrete (nominal) attributes, so the probability scores are grouped. The regression, (similar to partial least squares, and neural network algorithms) responds better to (numeric) continuous attributes, so the probability scores are more linear. Several authors have explored the use of regression models in healthcare data analysis and found them to be very suitable (Topuz et al., 2018; Rasella et al., 2014).

Analytical Tools

A plethora of analytical tools currently exist. The most common include R, Python, Hadoop, MongoDB (Saini & Kohli, 2018; Das, Pandey & Saxena, 2017), Stata, SPSS, Revolution Analytics, Alpine Data lab, and SAS amongst others. According to Acock (2005), SAS programs are versatile in the breadth and depth of their capability for data analysis and management although they have steep learning curves. While SPSS and Stata are easier to learn, they are less capable for analytics hence Acock concluded that SAS was best for power users. The SAS suites are offered to educational institutions for free. A study based on a sample of 1,139 articles drawn from three journals found that SAS was used in 42.6 percent of data analyses in health service research (Dembe, Patridge & Geist, 2011). Available tools including SAS suite provide incentives for researchers to mine data; develop actionable models; and help defuse the myth behind current *Big Data* conundrum.

It is common knowledge that data analytic techniques can improve healthcare practices and medical efficiency provided they are expertly applied (Murdoch & Detsky, 2013; Koh & Tan, 2005). What is not known is the preparedness of healthcare organizations to apply analytic tools to address business challenges. Another potential issue is users' ability to correctly pair analytical tools to business questions or challenge. This paper aims to demonstrate the use of SAS analytical techniques including DT and R to address a business challenge in diabetes healthcare.

2. DIABETES HEALTHCARE MANAGEMENT

Diabetes Healthcare

The American Diabetes Association (2018) estimated that the total costs of diagnosed diabetes rose by 26 percent from $245 billion in 2012 to $327 billion in 2017. The report also showed that people with diagnosed diabetes incur $16,752 average medical expenditures per year. About $9,601 of this is directly attributed to diabetes. As a result, diagnosed diabetics have medical expenditures approximately 2.3 times higher than what expenditures would be in the absence of diabetes. Care for people diagnosed

with diabetes account for 25% health care dollars in the U.S., and more than half of that expenditure is directly attributable to diabetes.

Maguire and Dhar (2012) found that the top 10% of newly diagnosed type II diabetes patients account for 68% of healthcare utilization. The global health and economic cost is enormous with type II diabetes accounting for 90-95% of all diagnosed cases of diabetes (CDC, 2017). Diagnosed diabetes account for more than 20% of health care spending in the United States (US). Although it may be underreported, diabetes was found to be the seventh leading cause of death in the United States in 2013. It leads to microvascular and macrovascular complications which result in enormous morbidity, disability and mortality (Young et al., 2008); it is also the leading cause of kidney failure, lower-limb amputations, and adult-onset blindness (CDC, 2017).

Given the enormous costs, limited socioeconomic resources, and healthcare organizations' limited ability to invest in disease management initiatives for high-risk diabetic patients, it is important to develop models that predict which patients are at highest risk of adverse medical outcomes. Big data analytics can enhance healthcare quality per unit of spend through optimum utilization of available healthcare data (Kumari & Rani, 2018) existing in several silos including claims, clinical, geomapping, condition, and pharmacy amongst others.

Business Challenge and Goals

Patients with diabetes either don't make enough insulin (type I diabetes) or can't use insulin properly (type II diabetes). Insulin is needed to allow blood sugar (glucose) needed for energy to enter body cells; however, when the body is inadequate in insulin or unable to use it effectively, blood sugar builds up and can lead to several complications (CDC, 2017). The business challenge is to move uncontrolled type II diabetic patient (H1AC > 7) to a controlled state (H1AC < 7). In order to accomplish this, we will (a) develop models and determine the factors associated with a patient designated as controlled; (b) we will score new data to predict the probability of a patient being controlled.

The key opportunity in deploying advanced analytics is the insight it provides in developing interventional approach toward the patient care optimization. Traditionally, management of the diabetic population health relied on arcane reactive approaches that have seen the economic costs and management of diabetics increase every decade. For example, if we have an insurance provider that wants to minimize the dollar risk for the covered diabetic patients; however, once the patient is admitted to the ER or hospital due to conditions that require immediate medical care, it's basically too late – as the treatment must occur and the insurance company must cover their agreed upon portion. Basically, the insurance provider is being descriptive or "looking in the past". However, what if the insurance provider was able to establish specific drivers and a probability for why the patient may have to visit the ER *before* the event actually occurred? Results can inform interventional programs that could save the insurance company dollars at risk for that patient. This method is regarded as being prescriptive or "looking ahead of the curve". This example is the new perspective of the insurance provider.

From the patients' perspective: Suppose drivers can be established for what is causing a patient to be uncontrolled with regards to their diabetic management and the probability of such an event is obtained; this exercise helps the patient to be more proactive regarding their health. Perhaps the insurance provider can establish health coach programs for these patients that are deemed to have a high probability of being uncontrolled and reach out to them – as this would not only improve the patient's care but also minimize the dollars at risk for the insurance provider.

Despite the promises of data mining in healthcare however, there are several challenges that often need to be overcome. For instance, data often exist in dissimilar technology platforms; inferring knowledge from complex heterogeneous patients and data sources may be quite difficult and complicated especially for the untrained, so also is the ability to leverage the patient/data correlations in longitudinal records. This may further exacerbate the already computationally difficult task of analyzing clinical data.

Objectives

This paper demonstrates the use of SAS suite and data mining steps involved in leveraging massive data sets in providing timely patient intervention and personalized care that could benefit components of a healthcare system including provider, payer, patient and management. Srinivas et al. (2010) decried the lack of effective analysis tools to discover hidden relationships and trends in data. Through the use of SAS analytical software, this article introduces the healthcare stakeholder to the specific use of machine learning tools and the possible challenges and/or

techniques associated with using these tools in the healthcare domain.

3. DATA SOURCES

Available Data, Sources and Preparation

Advanced analytics cannot be performed, until data sources and attributes are identified, cleaned and prepared for analyses. Three distinct data sources were identified for this article as follows: (a) Clinical table containing lab results at the patients level (from various hospital labs in a .XLSX format); (b) Demographic table containing information such as the patients' age and gender (from the insurance provider in a .CSV delimited format); (c) Geo-mapping table containing information for mapping such as city and state (from the insurance provider in a .TAB delimited format).

4. SAS STUDIO ANALYTICAL DATA PREPARATION PROCESS

The analytical techniques and tools used in this case study include the following: SAS® 9.4 (BASE); SAS® Studio 3.71; SAS® Enterprise Miner (EM) 14.3. The analytical data preparation process can be very time consuming and is essential in ensuring the analytics conducted is not only accurate but answers the specific business question desired. SAS Studio is a point-and-click, menu- and wizard-driven tool that empowers users to analyze data and publish results.

SAS Studio Data Preparation Flow

This flow is divided into 7 sections (see Figure 2 in the Appendix) as described below: We created a folder and library in SAS studio (this is basically a Libname statement); and imported the three different files (excel, csv, tab format) to create respective SAS datasets. The three datasets were merged, using patient_ID and performing an inner join on the clinical dataset. Overall, we have approximately seventy (70) thousand unique patient records and approximately 130 variables. Based on industry standard for diabetes management, we used the clinical measure of Hemoglobin A1c for the business rule creation of the Target – what we used as our Y for the predictive model of a controlled diabetic. A simple SAS program was written for this task:

```
if HA1c < 7 Then Diabetes_Controlled = 1;
Else 0.
```

Figure 3: SAS code for creating the target

We performed a one-way frequency exploration on the newly created variable to assess the theoretical soundness for predictive modeling. Based on experience, the best practice is to have 80% 0's and 20% 1's for a binary target (Y). Our results yielded about 68% 0's and 32% 1's – which is deemed acceptable. We would not wish to have a 50%/50% split or more 1's than 0's as that would not theoretically ensure a strong predictive model. The goal of the modeling process is to establish the drivers that increase the percentage of 1's (uncontrolled) and optimize those results.

Subsequently, we created a 10% random sample from the final dataset with the target. This is basically a litmus test that will be used for scoring the new model at the end of the EM flow. Since this 10% dataset will have the actual targets, we can immediately inspect if the model was able to predict those records accurately - specifically the 1's. This is a very important step in the data mining process before scoring a new dataset ("blind file") where no target exists. The final piece of the flow de-duplicates those 10% records from the final dataset that will be brought into EM for data mining & predictive modeling. This creates 90% of the raw datasets that will be brought into the EM for modeling. A simple SAS program to execute this is shown below. Once preparation processes concluded in SAS Studio, data was exported into the EM.

```
data health.diabetes_to_model_90pct;
merge health.diabetes_merged (in=a)
health.diabetes_to_be_scored_10pct (in=b);
if a and b then delete;
run;
```

Figure 4: SAS code for deduplication

5. DATA MINING & PREDICTIVE ANALYTICS

SAS EM and the SEMMA Process

The SAS Enterprise Miner is a machine learning tool that helps to streamline the data mining process for the user to create accurate predictive and descriptive analytical models using vast amounts of data. The EM functionality depends on the SEMMA process. SEMMA is an acronym for Sample, Explore, Modify, Model, and Assess - all pertain to conducting data mining and predictive modeling tasks. The process creates a specific

step-by-step strategy for executing this analytical exercise. It also allows for several check points along the way in case certain steps need to be modified or adjusted.

Adjusting the metadata for the input data node

There are four (4) roles of concern: (a) ID, (b) Target, (c) Input, (d) Rejected. Each role have different levels that were addressed including binary, interval, nominal, ordinal and unary.

Data Exploration to assess missing values

Certain models like regression and neural networks cannot have missing values for any variable otherwise those records will be deleted and made unavailable for modeling. Therefore, if any of these models are used, it is important that the practitioner make the necessary modifications by first exploring the data using StatExplore node in EM. Henceforth, the impute node can be activated as explained later in this chapter (Note that the best practice is to partition data before imputation is performed). In addition, there are two other pieces of output that offer business value: (a) a bar chart showing only the correlations between the nominal (categorical) inputs against the target (Y = controlled) in descending order (most associated to least associated); (b) a bar chart showing the correlations between all of the variables (nominal and interval) against the target. Figure 5 (see Appendix) shows three pieces of information which provide business value:

First, the upper left window shows how all the categorical variables correlated with the target in descending order. We observed that the number of adverse events is highly associated with a controlled diabetic. The bottom window shows all the variables (categorical and continuous) and their association with the target in descending order. The upper right window shows descriptive statistics output, which allowed us to inspect one critical column – Missing. This statistic is essential for modeling, because if a variable is missing, then the entire record is omitted from the model. Therefore, possible imputation methods may need to be applied to the categorical and continuous variables, respectively.

Data Partitioning

An important step prior to model building is to divide the data into training and validation datasets; best practice suggests 70:30 respectively. Note that the dataset being partitioned is the 90% raw dataset that was previously created in the SAS Studio. Furthermore, data partitioning needs to account for the distribution of 0's and 1's within the target – this is called stratification. The software handles this automatically. This ensures that both datasets (training and validation) will have almost the exact percentage of 0's and 1's within the target. Figure 6 provides a summary statistics for class targets.

Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Controlled	0	0	42455	67.9900	
Controlled	1	1	19988	32.0100	

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Controlled	0	0	29719	67.9913	
Controlled	1	1	13991	32.0087	

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Controlled	0	0	12736	67.9870	
Controlled	1	1	5997	32.0130	

Figure 6: Summary Statistics for Class Targets

Data Imputation

There are many techniques to use for imputation. It is best practice to use the DT method for nominal variables (this is due to the fact that DTs handle missing values in the growth of the trees). For the interval variables, the median is recommended (Figure 7) because it is more robust and less sensitive to outliers than other measurement such as the mean.

Property	Value
General	
Node ID	Impt
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Tree
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Median
Default Target Method	None
Default Constant Value	

Figure 7: EM Property Selection for Imputation

For the categorical (nominal) variables, there are a variety of techniques to choose from. The Tree replaces missing class variable values with values

that are estimated by analyzing each input as a target. The remaining input and rejected variables are used as predictors. The imputed value for each input variable is based on the other input variables, hence this imputation technique is more accurate than simply using the variable *Mean* or *Median* to replace the missing tree values. For the continuous variables, there are also a variety of techniques to choose from. A preferred method is the distribution technique, which replaces values that are calculated based on the random percentiles of the variable's distribution. The assignment of values is based on the probability distribution of the non-missing observations. This imputation method typically does not change the distribution of the data very much. Another common technique is the *Median*, which replaces missing interval variable values with the 50th percentile. The *Median* is less sensitive to extreme values than the *Mean* or *Midrange*. Therefore, the *Median* is preferable when you want to impute missing values for variables that have skewed distributions because it allows you to select for each variable.

Data Transformation

An important modification technique that helps normalize the data is to adjust left or right skewness in a variable (the heaviness of the tail in terms of kurtosis). There are many mathematical techniques to do this adjustment. The best practice which we also use in this study is the Log(x).

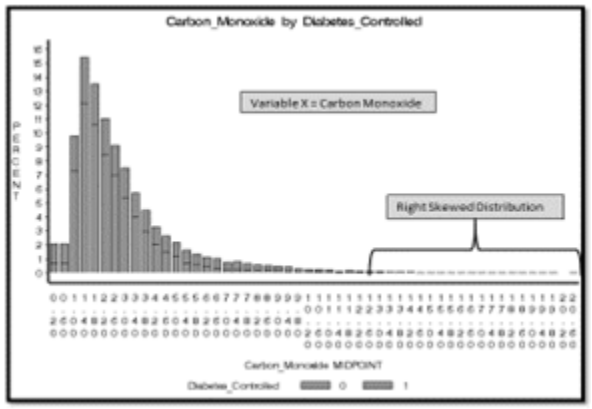


Figure 8a: Data Normalization Using the Log(x)

It is highly recommended to not use the overall method for variables, but rather, select specific transformation techniques for each variable. The overall method may tend to be dangerous as it is usually not the case that all categorical variables should have the same mathematical transformation applied, as in continuous variables. Figure 9 illustrates the second and recommended way, which selects specific techniques *a la carte* for the variables.

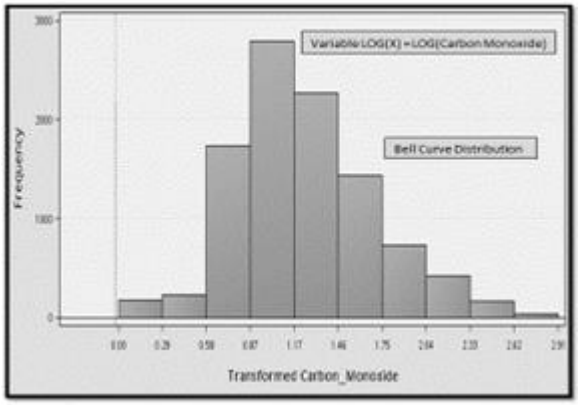


Figure 8b: Data Normalization Using the Log(x)

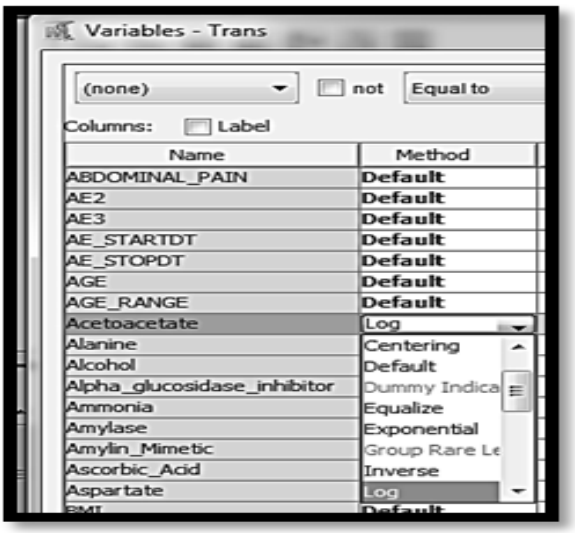


Figure 9: Variable Properties for Transformation Node

Building the models

It is important to note that although there are 14 modeling nodes (Figure 10), there are several different modeling techniques available within the property settings that result in much more than 14 algorithms. In this example, we will discuss two types of models (Decision Tree and Regression). Although 2 models are used in this study, there are several variations of techniques within the properties that can form different algorithms for our task of establishing drivers of a controlled diabetic.

Figure 10: SAS Enterprise Miner – SEMMA – Model Tab

For the Decision Tree node, we leave the default properties as is, and they can be viewed as "smart starts" tested and developed by SAS R&D for a good starting point in dial settings for the model. However for Regression node, and model, we adjust a couple of settings. The default type of regression is Logistic, which is what we need for our example, since we have a binary target of a controlled diabetic (0/1). We select the stepwise model selection method as it is a hybrid of a forward and backward method and conduct a variable selection-like process to choose the drivers of the model. In addition, the criterion we desire is the validation misclassification rate, since we wish to accurately classify a patient of being controlled, hence, we also wish to minimize the misclassification rate. Furthermore, we select the dataset that is designed to verify the model's development – the validation data set (see Figure 11).

Property	Value
General	
Node ID	Reg
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Misclassification
Use Selection Defaults	Yes

Figure 11: Property Values for Model Selection

Model Comparison
Once we build our models, an important aspect is to make a head-to-head comparison. The Model Comparison node is used for this assessment and can be found within the Assess tab of SAS EM as indicated below.

Figure 12: SAS Enterprise Miner – SEMMA – Assess Tab

SAS EM Model Comparison Flow
In order for EM to select the "best" model, we need to adjust some property settings. Figure 13 (see Appendix) shows the property setting adjustment for Model Comparison Flow.

Property	Value
General	
Node ID	MdlComp
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
Model Selection	
Selection Data	Default
Selection Statistic	Misclassification Rate
HP Selection Statistic	Default
SAS Viya Selection Statistic	
Selection Table	Validation
Selection Depth	10

Figure 13: Property Setting Adjustment for Model Comparison Flow

SAS Enterprise Miner – Model Comparison Statistics
The goal is to accurately classify a controlled diabetic from the model, hence, we select the misclassification rate, and also use the validation data set as the file to use for selection. The next step is very important, as we must assess the model from a theoretical standpoint, before proceeding to look at the individual model results. Although Figure 14 (see Appendix) indicates that both models are very similar in misclassification rate, the regression model (0.202477) is slightly better and is selected over DT (0.20253). Overall, the models show approximately 80% chance of accurately classifying a controlled diabetic patient.

6. RESULTS AND DISCUSSION

SAS Enterprise Miner – Model Comparison Results – Cumulative Lift
Another important theoretical checkpoint is the measurement of the cumulative lift vs. the depth on the graph (See Figure 15 in the Appendix). *What does this graph mean?* Let's imagine that we have a file of 1,000 patients where we desire to develop the probability of being a controlled diabetic. We then score that file with a winning model from EM, and sort the file in descending

order by deciles or depth. Now it makes sense in this graph that the cumulative lift line is decreasing down to 1 when the depth reaches 100% of the scored file. Also, the cumulative lift is highest at the smallest depth, say 5% of the scored file in descending order. *We now ask, what is Cumulative Lift?* The idea of lift refers to the predictive power or accuracy of the model at various deciles of a scored file. We would expect the predictive power to be very high when we are not deep into the scored file, because there is not a lot of variability among the observations. However, the deeper you get into that scored file, the more observations are found, which increases the variability, and makes the predictive power more difficult to achieve, until you finally score the entire file, and are left with no lift, or a lift of 1. Based on experience, we have found that the "sweet" spot of this graph is actually at the 20th depth or decile. Basically, assessing the cumulative lift value should be done at that point for true predictive power. Finally, it is best practice, to look at the 20th decile, a cumulative lift of at least 2 to deem a satisfactory model has been established. We notice, in our models, our cumulative lift is around 2.1, which meets our minimum criterion.

SAS EM – Model Comparison Results – Cumulative % Captured Response

To explain Figure 16 (see Appendix), let's use our example of scoring a patient file of 1,000 records to develop the probability of a controlled diabetic. Based on simple random chance theory, we would expect to obtain 20% of the events (Controlled Diabetics or 1's) as we score the top 20% of the 1,000 observation patient file. Also, we would also expect to obtain 40% of the events as we score the top 40% of the file, and so forth. Recall, from the previous graph, our best practice minimum cumulative lift is 2. If we notice in Figure 16, we are capturing about 40% of the controlled diabetics at the 20th decile, which actually is a cumulative lift of 2 (40% / 20th decile = 2). So think of the random chance theory as a diagonal line drawn from (0,0; 20,20; 40,40; etc.). Therefore, we desire to achieve, in any model, at least 40% of the events of interest at the 20th decile. We need to always be above that imaginary diagonal line for our model to be "good". Finally, this point is the true point of predictive power, and, we can state that if we are provided a file of 1,000 patients whose probability of being a controlled diabetic is desired, we can assume that within the top 20% (200 patients), we can obtain 40% of the controlled diabetics (80 patients). This becomes very powerful, if going deep into a scored file is costly from a health outreach and coaching perspective.

SAS Enterprise Miner – Model Comparison Results – Fit Statistics

The Output shown in Figure 17 (see Appendix) depicts a summary of stepwise selection of models and shows that the Regression model was selected over the DT model. In theory, since all the models are equal, anyone can be used for scoring. However, it is important to assess the drivers of the model used for scoring to ensure it makes sense to the business needs. In addition, choosing a model that is easy to explain from a business perspective is important. The Regression model below shows the factors driving a controlled diabetic below:

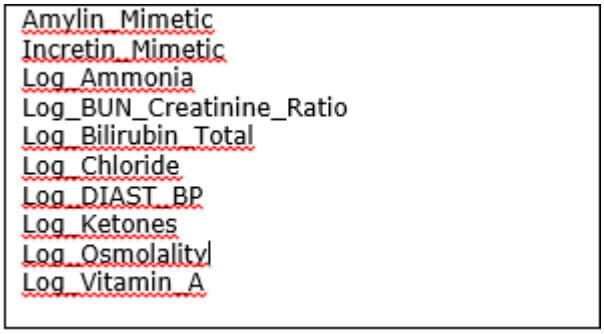


Figure 18: Drivers for controlled diabetic

Note that many of the main drivers selected by the model above, have been transformed using the log(x) technique.

SAS EM – Regression Scoring Results

Now that a solid predictive model has been established for the business question, the next step is to take the winning algorithm to score a patient file i.e. the remaining 10% raw data originally randomly selected and separated as previously indicated. Essentially, we are using the 10% datasets to verify the regression algorithm and develop the probability of controlled diabetic. The DT is a discrete algorithm, so the probability scores are grouped. However, the Regression, (similar to partial least squares, and neural network algorithms) is continuous, so the probability scores are more linear. The Figure below shows the first 10 patients sorted on their probability of being controlled (highest to lowest). This indicates that the model is very useful not only in assessing the distribution of probabilities but more importantly to decide on an appropriate cut-off for assigning which patients exhibit the probability of a given event (for example, being a controlled diabetic).

Patient_ID	Controlled	Probability for level 1 of Co...
1.303634424...	1.0	0.8170103596362283
6.927410991...	1.0	0.7924117380593262
9.876233050...	1.0	0.7897975478471009
7.797040216...	1.0	0.789405552385899
8.413553671...	1.0	0.7840471775353454
4.682905296...	1.0	0.7829350415489733
5.986654475...	1.0	0.7825617247187416
1.320028338...	1.0	0.7821573306515491
6.495120331...	0.0	0.7819917349673773
1.063950438...	0.0	0.7818783693274509

Figure 19: The first ten (10) Patients Sorted by Probability of Being controlled (highest to lowest)

Thus, any patient possessing a score greater than the cut-off score (often decided by the business) is deemed to be a "1"; otherwise, they are tagged as a "0". In this example, the probability of a patient being a controlled diabetic is calculated. It would be desirable for a majority of patients to be at the high end of the score spectrum, as that would mean lower risk of being uncontrolled and as a result would imply lower medical risk. The lower probability implies that the patients are uncontrolled and need better care management. The desire would be to move the population towards the higher end scores - being more controlled with their diabetes.

7. CONCLUSION

This paper demonstrates *big data* promise and potential in healthcare specifically using SAS analytical tool. Data preparation is critical before any data mining and predictive modeling can be executed. Figure 20 (see Appendix) shows the final flow of the algorithm described above, which were used in developing the model. In our specific example, the goal is to manage population health and gain better insight into the patients. Transitioning patients from being uncontrolled with their diabetes to a more controlled state is essential in minimizing risk and also optimizing care. The drivers of a controlled diabetic appear to be chemical factors each of which can be influenced with the right therapeutic intervention in uncontrolled diabetics. Given new patients records with corresponding diagnosis, data mining and analytic techniques demonstrated are able to determine patients at risk or the probabilities of a patient going/returning to the ER. This can cause enormous stress to the patients, health and socio-economic systems in terms of costs, morbidity and death. Hence early interventional initiatives can target such patients and potentially move them to a more desirable state resulting in better business decision making and dollar savings for health systems.

8 REFERENCES

Acock, A.C. (2005). SAS, Stata, SPSS: A Comparison. *Journal of Marriage and Family*. Retrieved September 21, 2005 from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1741-3737.2005.00196.x

Alkhatib, M.A., Talaei-Khoel, A & Ghapanchi, A. H. (2015). Analysis of Research in Healthcare Data Analytics. *Australasian Conference on Information Systems, 2015 Sydney.*

American Diabetes Association (2018). The Cost of Diabetes (March 22, 2018). Retrieved September 24, 2018 from: http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html

Bates, D.W., Saria, S., Ohno-Machado, L., Shah, A. & Escobar, G. (2014). Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. *Health Affairs 33, No. 7*(2014): 1123-1131

Center for Disease Control (CDC). 2017 https://www.cdc.gov/chronicdisease/resources/publications/aag/pdf/2016/diabetes-aag.pdf

Das, D., Pandey, R. & Saxena, A. (2017). Disease prediction using Hadoop with Python. Retrieved September 22, 2018 from: https://www.researchgate.net/publication/322714853_Disease_prediction_using_Hadoop_with_Python

Dembe, A.E., Patridge, J.S., & Geist, L.C. (2011). Statistical Software Applications Used in Health Services Research: Analysis of Published Studies in the US. *BMC Health Services Research 11*:252. Retrieved September 21, 2018 from: https://bmchealthservres.biomedcentral.com/track/pdf/10.1186/1472-6963-11-252

Hachesu, P.R., Ahmadi, M., Alizadeh, S. & Sadoughi, F. (2013). Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients. *Healthc Inform Res. 2013 Jun;19*(2):121-129. English. Published online June 30, 2013. https://doi.org/10.4258/hir.2013.19.2.121

Kajabadi A, Saraee M.H., Asgari S. (2009). Data mining cardiovascular risk factors; *Proceedings of International Conference on Application of Information and*

Communication Technologies; 2009 Oct 14-16; Baku, Azerbaijan. pp. 1-5

Kaur, A., & Arora, J. (2018). Heart Disease Prediction Using Data Mining Techniques: A Survey. Intl. J. of Advanced research in Computer Science vol 9(2). Retrieved September 23, 2018 from: http://www.ijarcs.info/index.php/Ijarcs/article/view/5872

Kaur, H. & Wasan, S.K. (2006). Empirical Study on Applications of Data Mining Techniques in Healthcare. *Journal of Computer Science 2 (2): 194-200, 2006*

Koh, H.C. & Tan, G. (2005). Data Mining Applications in Healthcare. *J Healthc Inf Manag. 2005 Spring; 19*(2):64-72.

Kumari, S. & Rani K.S. (2018), Big Data Analytics for Healthcare System (February 7, 2018). *IADS International Conference on Computing, Communications & Data Engineering* (CCODE) 7-8 February. SSRN: https://ssrn.com/abstract=3168338 or http://dx.doi.org/10.2139/ssrn.3168338

Maguire, J. & Dhar, V. (2012). Comparative effectiveness for oral anti-diabetic treatments among newly diagnosed type 2 diabetics: data-driven predictive analytics in healthcare. *Health Systems July 2013, Volume 2, Issue 2*, pp 73–92 Retrieved 12/31/2017 from: https://link.springer.com/article/10.1057/hs.2012.20

Murdoch, T.B. & Detsky, A.S. (2013). The Inevitable Application of Big Data to Health Care. *JAMA 2013; 309*(13):1351-1352. doi:10.1001/jama.2013.393

Raghupathi, W. & Raghupathi, V. (2014). Big Data Analytics in healthcare: promise and potential. *Health Information Science and Systems 2*(1) page 3.

Raghupathi W. (2010). "Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity." Edited by Kudyba S. Taylor & Francis; 2010:211–223

Rasella, D., Harhay, M.O., Pamponet, M.L., Aquino, R., & Barreto, M.L. (2014). Impact of primary health care on mortality from heart and cerebrovascular diseases in Brazil: a nationwide analysis of longitudinal data *BMJ 2014*;349 :g4014

Saini S., & Kohli S. (2018) Healthcare Data Analysis Using R and MongoDB. In: Aggarwal V., Bhatnagar V., Mishra D. (eds) Big Data Analytics. *Advances in Intelligent Systems and Computing*, vol 654. Springer, Singapore

Sitar-Taut, D.A. & Sitar-Taut, A.V. (2010). Overview on how data mining tools may support cardiovascular disease prediction. *J Appl Comput Sci 2010;4*(8):57–62.

Srinivas, K., Rani, B.K., & Govrdhan, A. (2010). Application for Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *Intl. J. on Computer Sc. and Engineering (IJCSE) vol 2* (2) pg. 250-255

Suryawanshi, R.D. & Thakore, D.M. (2012). Classification techniques of datamining to identify class of the text with fuzzy logic; *Proceedings of 2012 International Conference on Information and Computer Applications*; 2012 Feb 17-18; Hong Kong. pp. 263-267

Topuz, K., Zengul, F.D., Dag, A., Almehmi, A., & Yildirim, M.B. (2018). Predicting graft survival among kidney transplant recipients: A Bayesian decision support model. *Decision Support Systems, Volume 106*: Pages 97-109, https://doi.org/10.1016/j.dss.2017.12.004.

Young, B. A., Lin, E., Von Korff, M., Simon, G., Ciechanowski, P., Ludman, E. J. … Katon, W. J. (2008). Diabetes Complications Severity Index and Risk of Mortality, Hospitalization, and Healthcare Utilization. *The American Journal of Managed Care*, *14*(1), 15–23.

Appendices and Annexures

Patient_ID	AGE	GENDER	ADVERSE_EVENT START_DATE	ADVERSE_EVENT STOP_DATE	BUN_Creatinine Ratio	HEMOGLOBIN A1c	BMI	HYPERTENSION	RENAL DISEASE	City	State
85348444	73	F			2.10	4.24	23.12	0	0	Deer Lake	PENNSYLVANIA
507587021	82	F	April 04, 2010	April 12, 2010	53.58	11.49	24.82	0	0	Lake Park	NORTH CAROLINA
561197284	76	F			75.52	0.16	28.70	0	0	Altoona	ALABAMA
618214598	69	M	November 03, 2007	November 06, 2007	1.75	0.02	27.95	0	0	Lithonia	GEORGIA
743515800	90	F	July 25, 2011	August 03, 2011	49.33	6.68	25.47	0	0	Beersheba Springs	TENNESSEE
911507067	75	F	May 06, 2011	May 15, 2011	46.84	2.10	22.54	1	0	Madawaska	MAINE
1009556938	82	M	July 03, 2011	July 06, 2011	8.96	7.35	29.28	0	0	Herron	MONTANA
1319853858	75	F	August 02, 2007	August 11, 2007	18.28	20.59	23.70	0	0	Linden	MICHIGAN

Figure 1: A Portion of SAS Merged Dataset using Patient_ID as the Primary Key

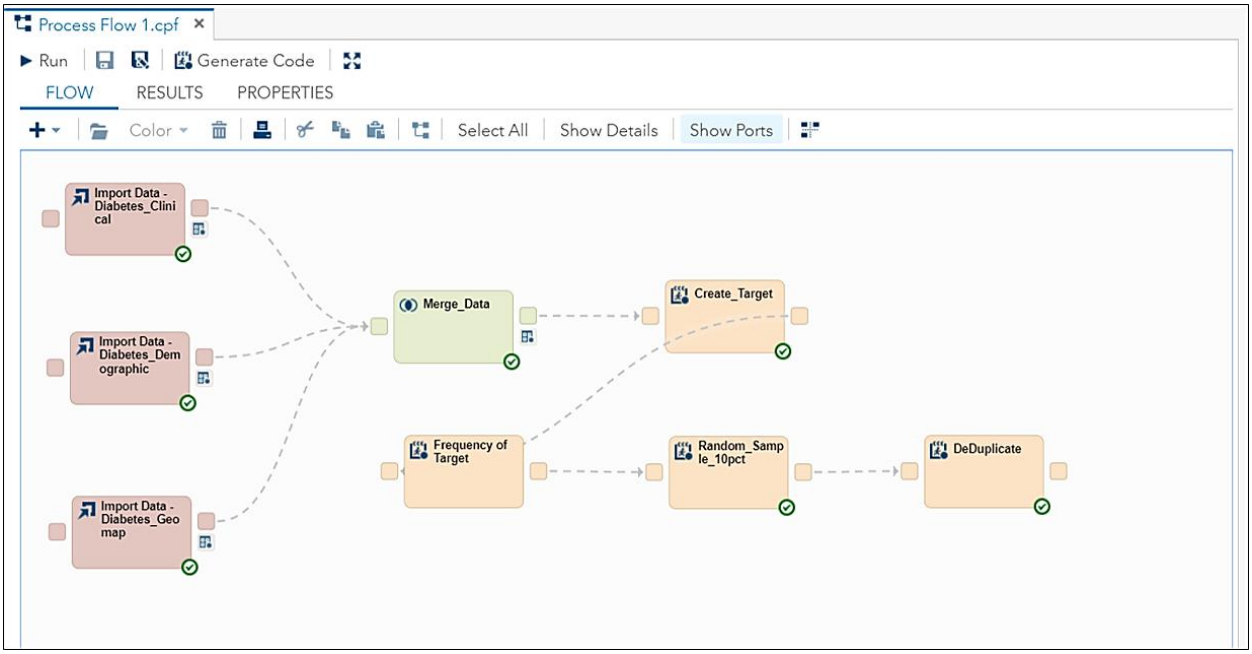


Figure 2: SAS Studio Data Preparation Flow

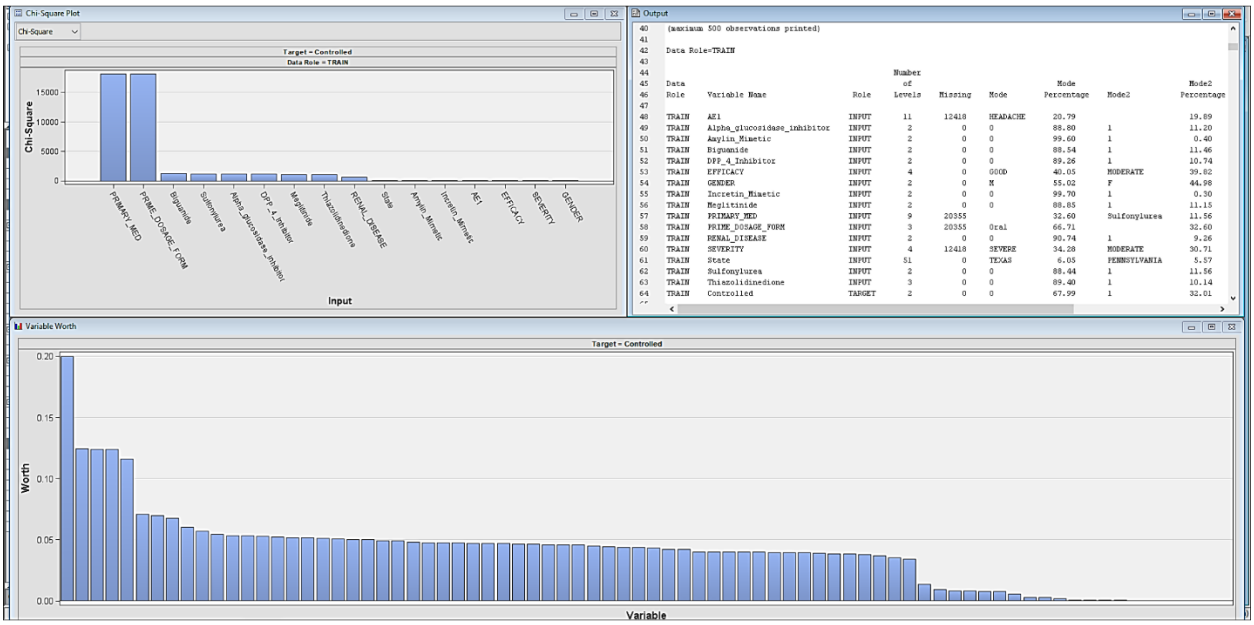
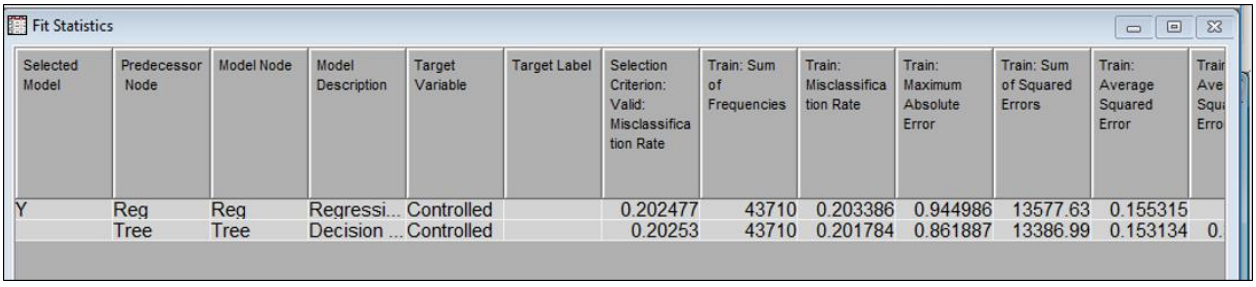


Figure 5: Result output to help Determine the Extent of Missing Data



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Trai Ave Squ Erro
Y	Reg	Reg	Regressi...	Controlled		0.202477	43710	0.203386	0.944986	13577.63	0.155315	
	Tree	Tree	Decision ...	Controlled		0.20253	43710	0.201784	0.861887	13386.99	0.153134	0.

Figure 14: Model Comparison Flow

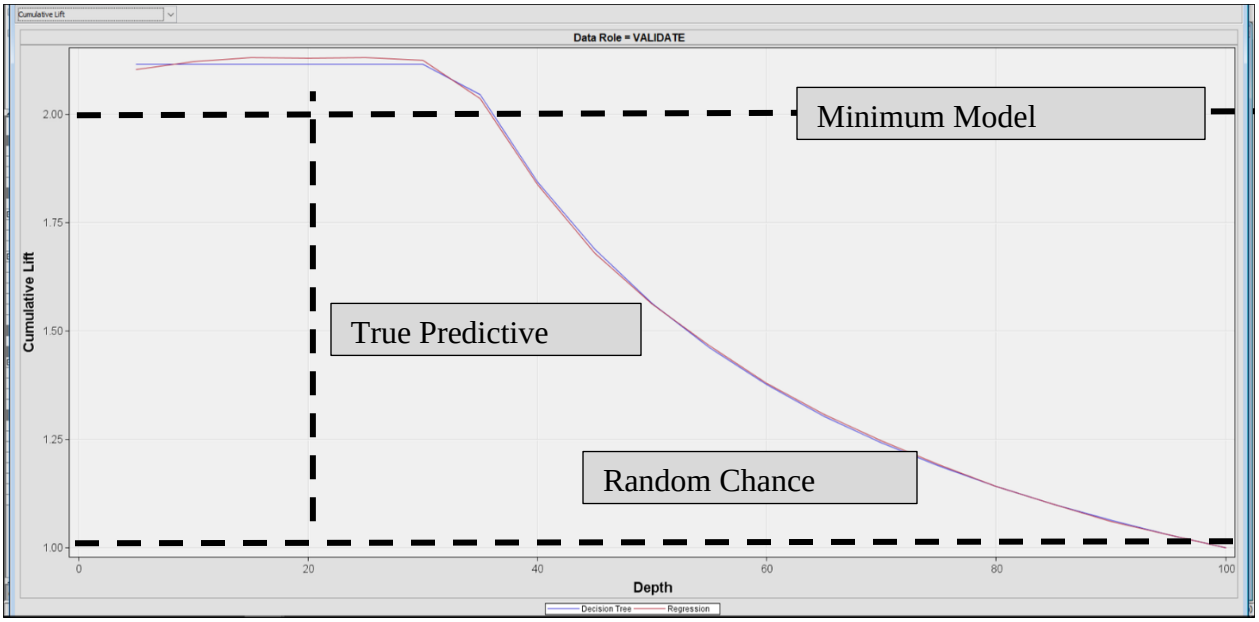


Figure 15: Model Comparison Results

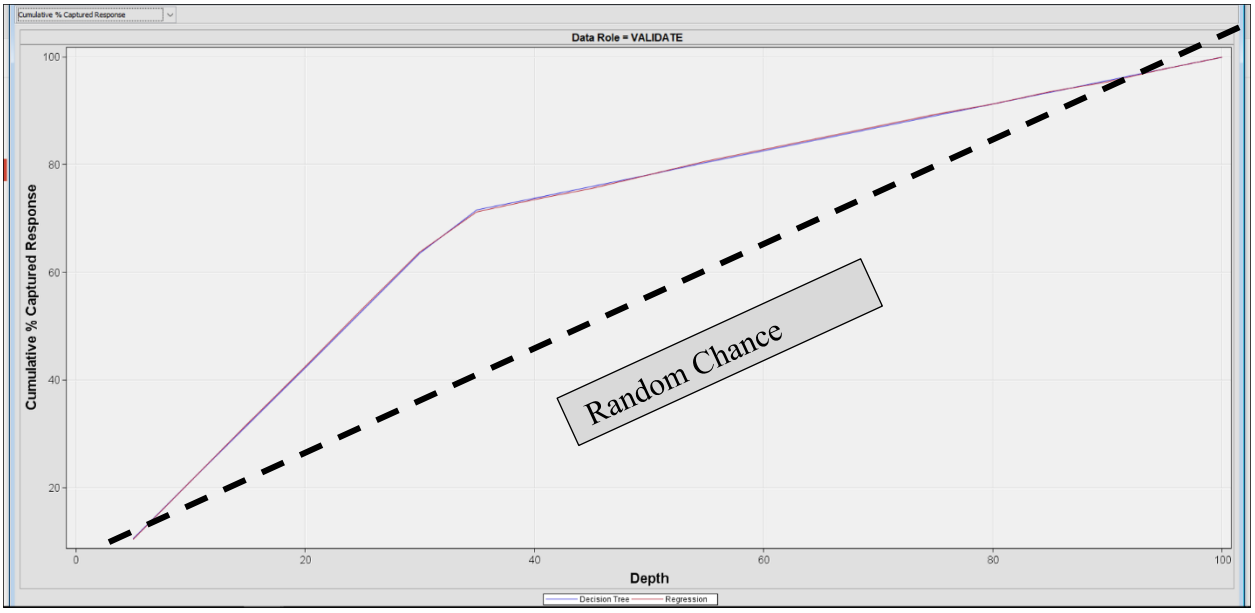


Figure 16: Model Comparison Results – Cumulative % Captured Response

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Misclassification Rate
1	LOG_Ketones	1	1	11907.8190		<.0001	0.2033
2	LOG_Bilirubin_Total	1	2	133.4161		<.0001	0.2033
3	LOG_Chloride	1	3	24.9012		<.0001	0.2032
4	LOG_DIAST_BP	1	4	21.3564		<.0001	0.2032
5	Amylin_Mimetic	1	5	19.1112		<.0001	0.2029
6	Incretin_Mimetic	1	6	19.4706		<.0001	0.2027
7	LOG_Osmolality	1	7	16.6654		<.0001	0.2026
8	LOG_Ammonia	1	8	14.6214		0.0001	0.2026
9	LOG_Vitamin_A	1	9	11.4710		0.0007	0.2026
10	LOG_BUN_Creatinine_Ratio	1	10	10.7005		0.0011	0.2025
11	LOG_Urine_PH	1	11	9.5332		0.0020	0.2026
12	LOG_Carbon_Monoxide	1	12	8.7770		0.0031	0.2026
13	LOG_Cholesterol	1	13	7.8737		0.0050	0.2027
14	LOG_Pyruvic_Acid	1	14	7.0667		0.0079	0.2027
15	LOG_Prostate_Specific_Antigen	1	15	6.1928		0.0128	0.2027
16	LOG_Blood_Volume	1	16	4.8873		0.0271	0.2028
17	LOG_Lactate	1	17	4.1612		0.0414	0.2029
18	LOG_MCV	1	18	3.9101		0.0480	0.2029

The selected model, based on the misclassification rate for the validation data, is the model trained in Step 10. It consists of the following effects:

Intercept Amylin_Mimetic Incretin_Mimetic LOG_Ammonia LOG_BUN_Creatinine_Ratio LOG_Bilirubin_Total LOG_Chloride LOG_DIAST_BP LOG_Ketones LOG_Osmolality LOG_Vitamin_A

Figure 17: Summary of Stepwise Selection

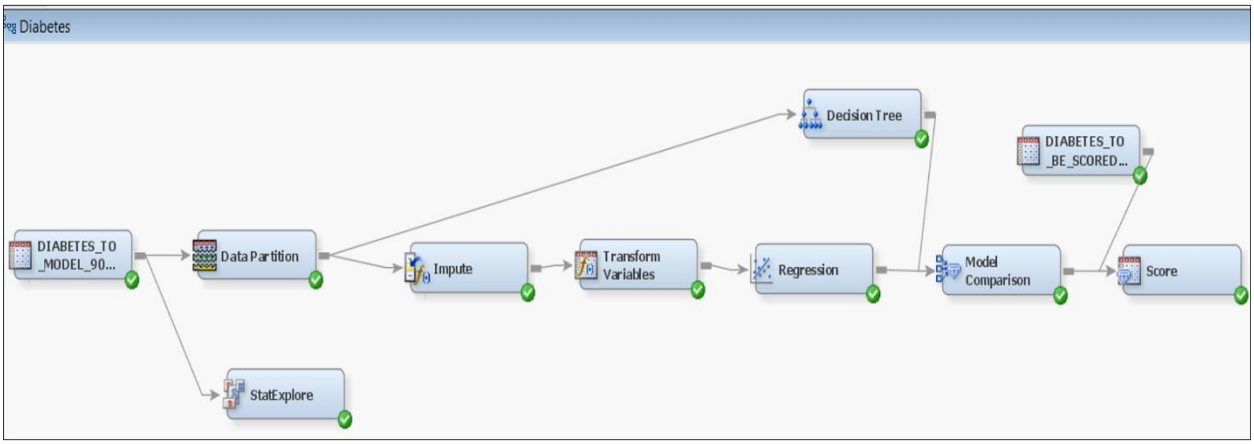


Figure 20: The Final Model Flow

Aligning the Factors of Skill, Technology and Task to Influence the Continued Use of Mobile Payment Users in China

Charlie Chen
chench@appstate.edu

Hoon Seok Choi
choihs@appstate.edu

Jason (Jie) Xiong
xiongjj@appsatete.edu

Department of CIS and SCM
Appalachian State University
Boone, NC 28608 USA

Yong Tang
tangyong@uestc.edu.cn
School of Computer Science and Engineering
University of Electronic Science and Technology of China

Abstract

Severe competition in the mobile payment market in China has resulted in user attrition issue. An effective business strategy to cope with high churn rate can arguably increase revenues and market share of mobile payment app and service providers. The present study adopts flow and task-technology fit (TTF) theories to investigate key factors contributing to the loyalty of mobile payment services in China as it has the largest mobile payment users in the world. A survey was conducted with 228 mobile payment users in China. An analysis of the collected data shows that users are more likely to perceive mobile payment services useful, satisfied, and willing to continue to use when the degree of alignment or fit among task, skill, and technology is improved. These findings offer theoretical and practical insights into lower user attrition rate, and improving user loyalty for mobile payment app and service providers.

Keywords: Mobile Payment Service, Task-Technology Fit Theory, Flow Theory, Loyalty, Perceived Usefulness

1. INTRODUCTION

Although mobile payment adoption rate is relatively low around the world, it is highly successful in China. Mobile payment services are ubiquitous and are becoming a dominant payment method in China (Wallis, 2015). More than 90% of 1.4 billion Chinese users are currently using smart phones to surf on the Internet. About 50% of them are active users of

mobile payment services. According to Wall Street Journal, from 2015-2016, mobile payment more than quadrupled, while in the United States the growth rate was only 40% (Abkowitz, 2018). The rapid adoption of mobile payment service has enabled China to become a leading cashless society (@WhartonKnows, 2018; Abkowitz, 2018). There are two main mobile payment service providers, Tencent, which owns WeChat Pay, and Ant Financial, which owns Alipay (Y. Wang, 2018). These two apps account for the majority of the mobile payment industry in China. The Economist newspaper has identified that financial technology, or fintech, including Alipay and WeChat pay, has revolutionized the behavior of Chinese consumers (Solutions, 2018).

The domination of the mobile payment market by two major local players (WeChat and Alipay) in China has created a market, where many domestic and international service providers compete for customer retention (e.g. Apple Pay, Xiaomi Pay, and Huawei Pay). However, there are several issues related to customer retention. First, users have trouble deciding which mobile payment app to use as too many options are available. Second, mobile payment apps are often not integrated well with various mobile payment services (e.g. money transfer, mobile commerce, renting a bike, paying for food, tickets and utilities). Third, the switching cost from one mobile payment service provider to another is low for users due to the competitive market. As such, users often do not care which digital wallet app to use as long as the purpose of each transactional services can be fulfilled. These reasons resulted in high user attrition in the market and thus, user retention is challenging for the mobile payment companies.

Current literature on mobile phone technologies primarily focuses on the antecedents for behavioral adoption intention of mobile phones, using technology adoption theory and economic consumption analysis (Zhang, 2017). Key factors identified in the literature include use expectancy, intrinsic motivation, requisite knowledge (K. H. Wang, Chen, & Chen, 2017), demographics, trust, performance expectancy, technology characteristics and task technology fit (Oliveira, Faria, Thomas, & Popovič, 2014). A growing number of scholars are exploring factors that cause the mobile divide between age, gender, and ethnic groups (Zhang, 2017). Some other scholars adopt the social network theory to explore the usage pattern of mobile applications. These studies identify key factors, such as social influence, perceived ease of use, perceived enjoyment and perceived interactivity (Lee, Park, Cho, & Jin, 2017). The factors identified in the previous study have been contributing to the maturity and development of smart phone adoption. However, there is little literature that is related to the primary factors of mobile payment customer retention.

Our study adopts the flow and task-technology fit (TTF) theories to examine key factors contributing to the loyalty of mobile payment services. Flow theory asserts that user satisfaction or ultimate pleasure relies on a good fit between skill and the challenge level of tasks (Csikszentmihalyi, 1990). Completing a mobile payment service (task) requires different skill sets (e.g. entering bank information, navigation, picture taking, scanning, and user interface). TTF theory extends our understanding of the fit between task and technology because mobile payment services are supported with various technologies (e.g. cell phone functions, features, bandwidth, mobile banking, and personalized marketing tools). According to TTF theory, information technology is more likely to have a positive impact on user performance and to be used if IT capabilities match the tasks performed by users (Goodhue, 1995). This implies that the fit should complement the dimensions of task and technology. Therefore, when the degree of fit among all three dimensions – task, skill and technology - increases, users are more likely to perceive mobile payment useful and satisfactory, and consequently, and to continue to use it. Our primary research question is about the continued use of mobile payments.

Based on the theoretical foundations, this study examines; (1) the interactions among the three key factors (i.e., skill, task-technology fit, and machine interaction), (2) how these factors individually and jointly influence user's perceived usefulness, (3) whether perceived usefulness can directly affect the flow experiences of mobile payment users, and (4) how to improve the user loyalty of users to mobile payment services via the increase of user satisfaction.

The remainder of this paper is organized as follows. *Conceptual Formation* discusses relevant literature to the aforementioned theories and major hypotheses of this study. *Research Methodology* introduces research design, data collection, and analysis method. *Hypothesis Test* presents results of the hypothesis test. *Theoretical and Practical Implications* discusses academic and practical contributions of this study. Finally, *Limitations and Future Research* discusses about the limitations of this study and

suggestions for future studies on mobile payment.

2. CONCEPTUAL FORMATION

Requisite skills to perform mobile payment

Flow theory asserts that a person can enter into a mental state of full immersion, involvement and enjoyment when their perceived challenges of the task matches with their perceived skills to achieve a clear set of goals or progress (Tse, 2018). Users engaging in mobile payment services often have a clear goal of completing a transaction with others (e.g. money transfer), vendor (e.g. ordering a product or a service), or government (e.g. filing income tax or paying utility bill). When users cannot complete the goal of a transaction (e.g. using smart phone to order a food delivery), they will worry and experience anxiety and arousal. On the other hand, they can feel apathy or boredom if completing a transaction (e.g. money transfer) is too easy. Therefore, it is imperative that users have a right requisite skill set of using mobile payment services in order to perform their transactional tasks.

Transactional tasks vary in complexity and require corresponding skill set in order to achieve a transactional goal. For instance, using a smartphone to rent a bike requires that a user to (1) download a bike rental app or install a mini app within mobile payment app, (2) scan personal ID to activate a bike rental account, (3) connect the bike account to bank account or smart phone's payment system for verification, (4) use the bike rental app to locate an available bike, (5) scan a barcode on the bike to check out the bike, and (6) file a claim if the rented bike is broken. Users will experience anxiety and will not be able to complete the goal of checking out bike if they do not possess the right skill set. In contrast, paying to a convenience store for a beverage requires that users (1) open up mobile payment app, (2) have the vendor scan the QR code on the mobile app, and (3) receive a transactional receipt. Some users will feel that completing such a simple task is not challenging. Users who have a right skill set to accomplish a transactional task are more likely to have positive influence on the increase of perceived task-technology fit. Thus, we propose:

H1: The right skill set to use mobile payment services has positive effect on the increase of perceived task-technology fit

Users tend to have positive experiences with mobile payment services if they have acquired the right skill sets to achieve transactional goals (e.g. mobile commerce, mobile top-up, pay credit card, buy lottery, money transfer, pay for leisure services, and overseas transfer). When users have positive experiences with new technology, they tend to have more positive perceptions about their usefulness (Ritu & Jayesh, 1998). Such a correlation is also likely to be present for the adoption of mobile payment services. When users with right skill sets to use mobile payment services and achieve transactional goals, they could become enthusiastic at communicating their usefulness to others. Therefore, users with right skills to use mobile payment are more likely to perceive it useful. Thus, we propose

H2: The right skill set to use mobile payment services has positive effect on the increase of perceived usefulness

Effective system use is comprised of three elements: user's competencies, the purpose of systems, and task characteristics (Andrew, 2013). However, when there is a mismatch among the three elements, users will have poor machine interaction experiences. A positive machine interaction indicates that a high degree of fit between individuals, tasks and technologies is achieved, which improves IS-enabled task performance (Goodhue, 1995). In face of poor machine interaction, users often try to compensate for the limitations of technology in order to achieve successful outcomes in using the technology (Serrano & Karahanna, 2016).

Mobile payment services also involve user's skill sets, the purpose of transactions, and task characteristics. In order to wire money to a friend, for instance, a user needs to open a mobile payment app, enter his friend's phone number, and follow the wiring instructions. To order a food delivery service, a user needs to learn how to install the food delivery app, connect mobile phone information to the app, enter delivery address, request a receipt, submit an order, track the order, and communicate with the deliver about the specific location. As task complexity increases, users need to improve their corresponding competencies of using varying mobile payment services in order to achieve different transactional goals. The variations of user's competence in using different mobile payment services are technology-specific user characteristics (Upadhyay, 2016). When users have positive interactions with mobile payment services, they are more likely to have (1) a high perceived task-technology fit and (2) a high perceived usefulness. Thus, we propose:

H3: Positive user interactions with mobile payment applications have positive effect on the increase of perceived task-technology fit

H4: Positive user interactions with mobile payment applications have positive effect on the increase of perceived usefulness

Achieving a fit between task and technology can lead to the effective use of information systems. A high degree of task technology fit (TTF) can be ensured with three steps of test: (1) clarifying distinct tasks to be completed, (2) adopting ideal information systems to support the completion of each task, and (3) testing the performance effects of task/technology alignments (Venkatraman, 1990). TTF theory is also applicable to understanding of mobile payment. First, mobile payment refers to a transaction task operated and performed from or via a mobile device. Second, completing a transaction anytime and anywhere involves mobility. Therefore, smart phone is an ideal information system to support mobile payment services. Third, users will find mobile payment is useful when a transactional task is successfully completed. Each mobile payment task dimension has a prescribed, best-fit technology dimension associated with mobile payment services. Increasing TTF can potentially increase the perceived usefulness of mobile payment services. Thus, we propose:

H5: A user's perceived task-technology fit of using mobile payment services has positive effect on the increase of perceived usefulness

When users have high degree of perceived usefulness of an information system, they tend to have a high satisfaction. The positive correlation has been supported by different theories, such as information system success theory, technology acceptance model, and affinity theory (Xu, 2018). Moreover, the logical relationship is evident in varying information systems, such as e-learning systems (Marjanovic, 2016), customer relationship management (CRM) systems (Tung, Lee, Chen, & Hsu, 2009), and building information model (G. Wang, & Song, J. , 2017).

User satisfaction with mobile payment services should be susceptible to the positive influence of perceived usefulness. Many studies have found that perceived usefulness plays a critical role for user's intention to adopt varying mobile services, such as mobile payment systems (Kalinic & Marinkovic, 2016; Liébana-Cabanillas, 2018; Thakur, Angriawan, & Summey, 2016), mobile banking services (X. Luo, 2010), mobile financial services (Y.-K. Lee, 2012), mobile commerce (M. Khalifa, 2008), and mobile application (H. Verkasalo, 2010). The skill sets required to perform these mobile services are replicable in advanced mobile payment services with varying mobile commerce applications. Given the previous studies that have shown that a positive relationship between perceived usefulness and user satisfaction in the adoption of mobile payment services (X. Chen, & Li, S. , 2017), it is plausible that the relationship would remain constant in the mobile payment. This discussion introduce the following hypothesis:

H6: A user's high perceived usefulness of mobile payment services has positive effect on the increase of user satisfaction

User satisfaction is an important perquisite for continuance intention to use information systems. When users have high satisfaction with an information system, they tend to have high loyalty (Xu, 2018). Previous studies found the empirical evidences in a range of information system domains, such as clinical information systems (Hadji, 2016), e-learning information systems (T. Lin, & Chen, C. , 2012), and social media (Kaewkitipong, 2016). The findings also corroborate with those of mobile payment (Cao, 2018; X. Chen, & Li, S. , 2017). Thus, we propose:

H7: A user's high user satisfaction with mobile payment services has positive effect on the increase of user stickiness with mobile payment services

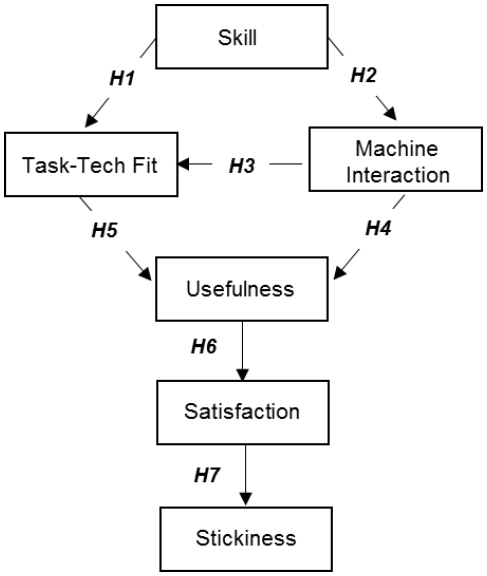


Figure 1. Research Model

The previous discussion leads to the development of the research model of this study (Figure 1).

3. RESEARCH METHODOLOGY

Survey Design

We designed a survey to measure constructs of the proposed hypotheses by employing established items in the previous studies (Please see Appendix 1). Questions for the constructs are in the 7-likert scale from 1 "strongly disagree" to 7 "strongly agree". We collected the data from WeChat users, which is one of the most popularly used mobile payment apps in China, to rule out possible differences among the apps that should affect the constructs.

Category	Group	Frequency	Portion
Gender	Males	119	52%
	Females	109	48%
Age (Years Old)	18-22	73	32%
	23-39	112	49%
	30-39	16	7%
	over 40	27	12%
Longevity of WeChat Use (Years)	< 0.5	5	2%
	0.5 – 1	16	7%
	1- 1.5	49	22%
	1.5 – 2	35	15%
	2 – 2.5	123	54%
Frequency of WeChat Use per Day (Times)	< 1	2	0.8%
	2 - 4	36	15.8%
	5 - 6	44	19.3%
	7 - 9	15	6.6%
	> 10	131	57.5%

Table 2. Profile of Respondents

Data Collection

We conducted an online survey for students in a Chinese university to collect data to test our proposed hypotheses. We had a total of 228 valid responses for analysis. The survey has a balanced ratio in gender while most respondents (81%) are in the age group between 18 and 39. This corresponds to the major user group of mobile payment services in China. In terms of longevity and frequency of WeChat Use, the majority of the respondents have used it for more than 1 year (91%) and more than 2 times per day (99.2%). Therefore, we can assume that the respondents are active users of WeChat with sufficient experience, who are adequate to provide reliable data for understanding mobile payment behaviors. Table 2 summarizes profiles of the respondents.

Reliability and Validity Tests

We conducted several reliability and validity tests to ensure adequacy of our research design. Before the tests, we performed Bartlett's test of sphericity and Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy statistics to examine whether the items measure distinct factors. All Bartlett's sphere values are significant ($p>0.01$), indicating that the correlation matrix does not have an identity matrix. All KMO values higher than 0.5, indicating that further factor analysis is adequate for the data (Kaiser, 1974).

All item loadings in factor analysis for indicator reliability are larger than 0.7 and therefore, all items for the constructs were included in our analysis. For testing internal consistency validity, Cronbach's α coefficients of the measurement are larger than 0.7, the acceptable threshold (Chin, 2010). To ensure convergent validity, we performed an analysis for average variance extracted (AVE) and composite reliability. The smallest AVE (0.521) is greater than 0.5, which is the recommended cut-off value (Fornell & Larcker, 1981; Hulland, 1999).

Cons.	CR	AVE	VIF
SK	0.896	0.613	1.88
MI	0.878	0.563	1.70
PU	0.865	0.521	1.56
PTTF	0.874	0.560	1.69
USF	0.864	0.525	1.58
USN	0.855	0.533	1.69

※ **CR:** Composite Reliability, **AVE:** Average Variance Extracted, **VIF:** Variance Inflation Factor, Square of AVE on the diagonal in bold

Table 3. Quality Indicators of Reliability and Validity

Likewise, all the values for composite reliability (c.f., the smallest is 0.85) are higher than the acceptable threshold, 0.7 (Fornell & Larcker, 1981). Table 3 summarizes the results of reliability and validity tests discussed above. Square root of the construct's AVE exceeds the correlations with other constructs, which ensures

discriminant validity of the measurement, as presented in Table 4. In addition, we estimated variance inflation factors (VIFs) for multicollinearity among constructs. The largest VIF is 5.43, which is significantly lower than the recommended cut-off of 10 (Chin, 2010). Therefore, the model does not have significant multicollinearity.

Cons	SK	MI	PU	PTTF	USF	USN
SK	**0.61**					
MI	0.45	**0.56**				
PU	0.45	0.43	**0.52**			
PTTF	0.27	0.32	0.33	**0.56**		
USF	0.39	0.27	0.23	0.46	**0.53**	
USN	0.02	0.04	0.06	0.17	0.31	**0.53**

Table 4. Correlations with Square Root of AVE on the Diagonal

4. HYPOTHESIS TEST

Structural Model

In order to test the proposed hypotheses, we adopted Structural Equation Modeling (SEM) with Partial Least Square (PLS). The estimation approach has advantages in that it can test multiple causal relationships (Henseler, Ringle, & Sinkovics, 2009) that our research model has. In addition, it is less subjects to the issues caused by scale of measurement and residual distribution (Fornell & Bookstein, 1982).

Analysis Results

Table 5 summarizes the PLS analysis results, including path coefficients and t-statistics for path significance.

Hypothesis	Coeff.	t-stat.	Test Result
H1: SK → PTTF	0.289	3.38**	*Supported*
H2: SK → PU	0.405	4.87**	*Supported*
H3: MI → PTTF	0.412	4.31**	*Supported*
H4: MI → PU	0.275	3.07**	*Supported*
H5: PTTF → PU	0.319	3.32**	*Supported*
H6: PU → USF	0.567	5.80**	*Supported*
H7: US → USN	0.719	6.02**	*Supported*

※ Significance: **p <0.01

Table 5. Results of Hypothesis Testing

All seven hypotheses are supported at the 99% confidence level. Hypothesis 1 is supported (β=0.289; p<0.01), suggesting that a right skill set to use mobile payment increases perceived task-technology fit of its users. The significant and positive relationship between skill and perceived task-technology fit (β=0.405; p<0.01) supports Hypothesis 2. This indicates that as mobile payment users are more skillful in using it, they perceive it more useful. Support of Hypothesis 3 (β=0.412; p<0.01) and Hypothesis 4 (β=0.275; p<0.01) suggests that as machine interaction, which presents interactivity between users and mobile payment technology, increases, the user perceive it is more suitable and useful to complete the payment task. The significant and positive relationship between perceived task-technology fit and perceive usefulness (β=0.275; p<0.01) indicates that mobile payment users who have a high level of task-technology fit would perceive it more useful, supporting Hypothesis 5. The perceive usefulness is found to have a positive relationship with user satisfaction (β=0.567; p<0.01), supporting Hypothesis 6. Lastly, the user satisfaction has a significant and positive relationship with user stickiness (β=0.719; p<0.01) and thus, Hypothesis 7 is supported.

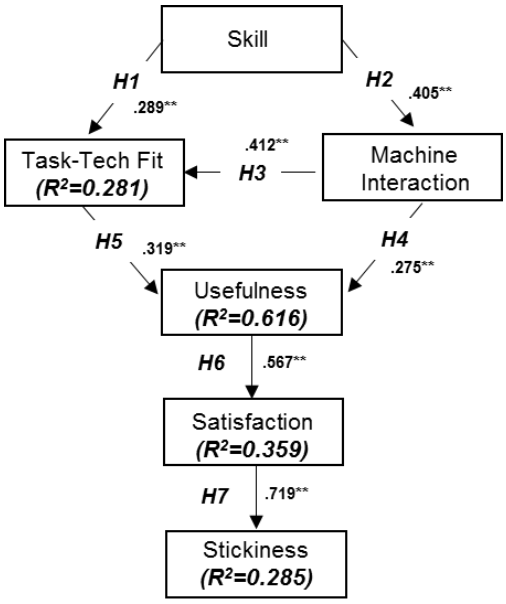


Figure 2. Research Model with Test Results

As illustrated in Figure 2, the two constructs Skill and Machine Interaction explain 28.1% of variation of Perceived Task-Technology Fit and 61.6% of variation of Usefulness, which does

35.9% of variation of Satisfaction. Lastly, Satisfaction is found to account for 28.5% of variation of Stickiness.

5. THEORETICAL AND PRACTICAL IMPLICARTIONS

Innovation in mobile payment is one of the emerging technologies that have the potential to change the dynamics of IT industry (Panetta, 2017). In China, although mobile payment has a relatively short history, its dramatic growth and saturation of the market have introduced numerous challenges to mobile payment companies, including retention of their customers. In the context of the Chinese mobile payment market, the findings of this study provide both theoretical and practical implications.

Grounded in flow and task-technology fit (TTF) theories, key factors contributing to the loyalty of mobile payment in China are identified. Support of Hypotheses 1 and 2 suggests appropriate skill sets of the mobile payment users positively affect the perceived task-technology fit as well as the perceived usefulness. These findings confirm that a fit between the skill sets and the level of difficulty is necessary to increase customer satisfactions (Goodhue, 1995). They also suggest that mobile payment providers need to consider strategies to equip users with proper skill sets to operate mobile payment services. For example, cultural differences, which is beyond language differences, may influence to proficiency about a technology use (Dai & Palvi, 2009; Harris, Rettie, & Cheung, 2005; Lu, Wei, Yu, & Liu, 2017). Shuter & Chattopadhyay (2014) shows that cultural value (e.g. vertical vs. horizontal individualism) affects the users' intention to engage in smart phone functions. Mobile payment app developers may want to understand the functional preference of users given their cultural background, and customize the design of mobile payment services. In addition, individual characteristics, such as professions, regions, and social influence (Bachfischer, Lawrence, & Steele, 2004; Kim, Mirusmonov, & Lee, 2010; Yang, Lu, Gupta, Cao, & Zhang, 2012), affect the perceived task technology fit. The sophistication level of smart phone technology needs to be tailed to the phone use based on users' characteristics (Serdarevic, Fazzino, MacLean, Rose, & Helzer, 2016).

Support of Hypotheses 3 and 4 indicates that the effective and efficient design of machine interaction have a positive impact on perceived task-technology fit and perceived usefulness. This further confirms that the IS-enabled task performance can be enhanced by the alignment of machine interaction and the task characteristics (Andrew, 2013). In practice, while complexity of mobile payment process may vary, the service providers need to be aware of the different competencies of using mobile payment services.

Support of Hypothesis 5 confirms that perceived TTF of using mobile payment has a positive effect on the increase of its perceived usefulness. This finding offers a practical implication that mobile payment service providers may consider collecting information about the perceived TTF and usefulness from users. For example, surveys or in-depth interviews may help develop future features of mobile payment apps. Listening to various customers may provide important, useful implications to improve their app design (Hsu & Liou, 2017).

Finally, support of Hypotheses 6 and 7 implies that high perceived usefulness of mobile payment has a positive effect on the increase of user satisfaction, and high user satisfaction with mobile payment services has a positive effect on the increase of user stickiness with mobile payment. As an important prerequisite for continuance intention to use, high perceived usefulness of mobile payment service is critical to the future growth of the services as well as the entire market. This finding corresponds to that of the previous mobile payment studies of X. Chen and Li (2017); Park, Jun, and Park (2017). Therefore, practitioners in the market need focus on improving user satisfaction and stickiness for their success.

6. LIMITATIONS AND FUTURE RESEARCH

There are several limitations of the research, as other studies. First, all the samples are students in one university in China. This may introduce a bias as most of the students might have similar experiences in mobile payment systems. Second, response biases, including acquiescence bias, demand bias, and social desirability bias, may exist during the survey data collection process, and their pontifical influence on our findings were not assessed in the study. Third, the survey limits mobile payment services to only one mobile payment app – WeChat. User behaviors in other mobile payment apps (e.g. Apple Pay, Alipay) could be different from those of WeChat, decreasing generalizability of our findings. Although it enabled a more rigorous research design, however, it did not allow considering behavioral difference across mobile payment

platforms. In the future, researchers may consider conducting a survey with diverse samples as well as having more samples from different locations in the world. They also may investigate perceptions and behaviors of other mobile payment apps such as Apple Pay and Alipay in the future.

7. REFERENCES

@WhartonKnows. (2018). The Mobile Payments Race: Why China Is Leading the Pack - for Now - Knowledge@Wharton. Retrieved from http://knowledge.wharton.upenn.edu/article/how-will-chinas-overseas-mobile-payment-systems-fare/

Abkowitz, A. (2018). The Cashless Society Has Arrived— Only It's in China. Retrieved from https://www.wsj.com/articles/chinas-mobile-payment-boom-changes-how-people-shop-borrow-even-panhandle-1515000570

Andrew, B., & Camille, G. . (2013). From Use to Effective Use: A Representation Theory Perspective. *Information Systems Research, 3*, 632.

Bachfischer, A., Lawrence, E., & Steele, R. (2004). *Towards understanding of factors influencing user acceptance of mobile payment systems.* Paper presented at the IADIS International Conference WWW/Internet.

Cao, X., Yu, L., Liu, Z., Adeel, L., & Gong, M. . (2018). Understanding mobile payment users' continuance intention: a trust transfer perspective. *Internet Research, 28*(2), 456-476.

Chen, X., & Li, S. . (2017). Understanding Continuance Intention of Mobile Payment Services: An Empirical Study. *Journal Of Computer Information Systems, 57*(4), 287-298.

Chen, X., & Li, S. (2017). Understanding continuance intention of mobile payment services: an empirical study. *Journal of Computer Information Systems, 57*(4), 287-298 %! Understanding continuance intention of mobile payment services: an empirical study %@ 0887-4417.

Chin, W. W. (2010). How to write up and report PLS analyses. In *Handbook of partial least squares* (pp. 655-690): Springer.

Csikszentmihalyi, M. (1990). Flow. The Psychology of Optimal Experience. New York (HarperPerennial) 1990.

Dai, H., & Palvi, P. C. (2009). Mobile commerce adoption in China and the United States: a cross-cultural study. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems, 40*(4), 43-61 %! Mobile commerce adoption in China and the United States: a cross-cultural study %@ 0095-0033.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Mis Quarterly*, 319-340.

Fornell, C., & Bookstein, F. L. (1982). Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory. *Journal of Marketing Research (JMR), 19*(4), 440-452.

Fornell, C., & Larcker, D. F. (1981). Structural Equation Models With Unobservable Variables and Measurement Error: Algebra and Statistics. *Journal of Marketing Research (JMR), 18*(3), 382-388.

Goodhue, D. L., & Thompson, R. L. . (1995). Task-Technology Fit and Individual Performance. *MIS Quarterly, 19*(2), 213-236.

H. Verkasalo, C. L.-N., F.J. Molina-Castillo, & H. Bouwman. (2010). Analysis of users and non-users of smartphone applications. *Telematics and Informatics, 27*(3), 242-255.

Hadji, B., & Degoulet, P. . (2016). Information system end-user satisfaction and continuance intention: A unified modeling approach. *Journal Of Biomedical Informatics*, 61185-61193.

Harris, P., Rettie, R., & Cheung, C. K. (2005). Adoption and usage of m-commerce: A cross-cultural comparison of Hong Kong and the United Kingdom. *Journal of Electronic Commerce Research, 6*(3), 210-224.

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In *New challenges to international marketing* (pp. 277-319): Emerald Group Publishing Limited.

Hsu, L.-C., & Liou, D.-K. (2017). Maintaining customer-brand relationships in the mobile

industry: the mediation effects of brand relationship benefits and quality. *International Journal of Mobile Communications, 15*(4), 388-413 %! Maintaining customer-brand relationships in the mobile industry: the mediation effects of brand relationship benefits and quality %@ 1470-1949X.

Hulland, J. (1999). Use of partial least squares (PLS) in strategic management research: A review of four recent studies. *Strategic management journal*, 195-204.

Kaewkitipong, L. l., Chen, C. C., & Ractham, P. . (2016). Using social media to enrich information systems field trip experiences: Students' satisfaction and continuance intentions. *Computers In Human Behavior*, 63256-63263.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31-36.

Kalinic, Z., & Marinkovic, V. (2016). Determinants of Users' Intention to Adopt M-Commerce: An Empirical Analysis. *Information Systems and e-Business Management, 14*(2), 367-387. doi:https://link.springer.com/journal/volumesAndIssues/10257

Kaur, P., Dhir, A., & Rajala, R. (2016). Assessing flow experience in social networking site based brand communities. *Computers in Human Behavior, 64*, 217-225.

Kim, C., Mirusmonov, M., & Lee, I. (2010). An empirical examination of factors influencing the intention to use mobile payment. *Computers in Human Behavior, 26*(3), 310-322 %! An empirical examination of factors influencing the intention to use mobile payment %@ 0747-5632.

Koufaris, M. (2002). Applying the technology acceptance model and flow theory to online consumer behavior. *Information Systems Research, 13*(2), 205-223.

Lee, S., Park, E.-A., Cho, M., & Jin, B. (2017). Factors affecting tablet computer users' intention to purchase mobile applications. *Social Behavior and Personality, 46*(1), 25-38.

Liébana-Cabanillas, F., Marinkovic, V., Ramos de Luna, I., & Kalinic, Z. . (2018). Predicting the determinants of mobile payment acceptance: A hybrid SEM-neural network approach. *Technological Forecasting & Social Change*, 129117-129130.

Lin, T.-C., & Huang, C.-C. (2008). Understanding knowledge management system usage antecedents: An integration of social cognitive theory and task technology fit. *Information & Management, 45*(6), 410-417.

Lin, T., & Chen, C. . (2012). Validating the Satisfaction and Continuance Intention of E-Learning Systems: Combining TAM and IS Success Models. *International Journal Of Distance Education Technologies, 10*(1), 44-54.

Lu, J., Wei, J., Yu, C.-s., & Liu, C. (2017). How do post-usage factors and espoused cultural values impact mobile payment continuation? *Behaviour & Information Technology, 36*(2), 140-164 %! How do post-usage factors and espoused cultural values impact mobile payment continuation? %@ 0144-0929X.

M. Khalifa, K. N. S. (2008). Explaining the adoption of transactional B2C mobile commerce. *Journal of Enterprise Information Management, 21*(2), 110-124.

Marjanovic, U., Delic, M., & Lalic, B. . (2016). Developing a Model to Assess the Success of E-Learning Systems: Evidence from a Manufacturing Company in Transitional Economy. *Information Systems And E-Business Management, 14*(2), 253-272.

Novak, T. P., Hoffman, D. L., & Yung, Y.-F. (2000). Measuring the customer experience in online environments: A structural modeling approach. *Marketing science, 19*(1), 22-42.

Oliveira, T., Faria, M., Thomas, M. A., & Popovič, A. (2014). Extending the understanding of mobile banking adoption: When UTAUT meets TTF and ITM. *International Journal of Information Management, 34*(5), 689-703.

Panetta, K. (2017). Top Trends in the Gartner Hype Cycle for Emerging Technologies, 2017. In: https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/.

Park, M., Jun, J., & Park, H. (2017). Understanding Mobile Payment Service Continuous Use Intention: An Expectation-

Confirmation Model and Inertia. *Quality Innovation Prosperity, 21*(3), 78.

Ritu, A., & Jayesh, P. (1998). A Conceptual and Operational Definition of Personal Innovativeness in the Domain of Information Technology. *Information Systems Research*(2), 204.

Serdarevic, M., Fazzino, T. L., MacLean, C. D., Rose, G. L., & Helzer, J. E. (2016). Recruiting 9126 Primary Care Patients by Telephone: Characteristics of Participants Reached on Landlines, Basic Cell Phones, and Smartphones. *Population health management, 19*(3), 212-215.

Serrano, C. I., & Karahanna, E. (2016). The Compensatory Interaction Between User Capabiities and Technology Capabilities in Influencing Task Performance: An Empirical Assessment in Telemedicine Consultations. *Mis Quarterly, 40*(3), 597-621.

Shuter, R., & Chattopadhyay, S. (2014). A cross-national study of cultural values and contextual norms of mobile phone activity. *Journal of Multicultural Discourses, 9*(1), 61-70.

Solutions, E. D. (2018). Plugging in: mobile payments in China. http://www.eiu.com/industry/article/915926475/plugging-in-mobile-payments-in-china/2017-11-13

Thakur, R., Angriawan, A., & Summey, J. H. (2016). Technological opinion leadership: The role of personal innovativeness, gadget love, and technological innovativeness. *Journal of Business Research, 69*, 2764-2773. doi:10.1016/j.jbusres.2015.11.012

Tse, D. K., Nakamura, J., Csikszentmihalyi, M., & Fung, H. H. . (2018). Teamwork and flow proneness mitigate the negative effect of excess challenge on flow state. *Journal Of Positive Psychology, 13*(3), 284-289.

Tung, F.-C., Lee, M. S., Chen, C.-C., & Hsu, Y.-S. (2009). An extension of financial cost and TAM model with IDT for exploring users' behavioral intentions to use the CRM information system. *Social Behavior and Personality: an international journal, 37*(5), 621-626.

Upadhyay, P., & Jahanyan, S. . (2016). Analyzing user perspective on the factors affecting use intention of mobile based transfer payment. *Internet Research, 26*(1), 38.

Venkatraman, N., and Prescott, J.E. . (1990). Enviornment-strategy coalignment: An empirical test of its performance implications. *Strategic Management Journal, 11*(1), 1-23.

Wallis, C. (2015). *Technomobility in China: Young migrant women and mobile phones*: NYU Press.

Wang, G., & Song, J. . (2017). Computers In Human Behavior. *The relation of perceived benefits and organizational supports to user satisfaction with building information model (BIM)*, 68493-68500.

Wang, K. H., Chen, G., & Chen, H.-G. (2017). A model of technology adoption by older adults. *Social Behavior and Personality: an international journal, 45*(4), 563-572.

Wang, Y. (2018). China Tightens Regulation Over Mobile Payment Apps -- What's Next For Tencent and Ant Financial? https://www.forbes.com/sites/ywang/2018/01/03/china-tightens-regulation-over-mobile-payment-apps-whats-next-for-tencent-and-ant-financial/

Wu, J.-H., Wang, S.-C., & Tsai, H.-H. (2010). Falling in love with online games: The uses and gratifications perspective. *Computers in Human Behavior, 26*(6), 1862-1871.

X. Luo, H. L., J. Zhang, J.P. Shim. (2010). Examining multi-dimensional trust and multi-faceted risk in initial acceptance of emerging technologies: An empirical study of mobile banking services. *Decision Support Systems, 49*, 222-234.

Xu, F., & Du, J. T. . (2018). Factors influencing users' satisfaction and loyalty to digital libraries in Chinese universities. *Computers In Human Behavior*, 8364-8372.

Y.-K. Lee, J.-H. P., N. Chung, & A. Blakeney. (2012). A unified perspective on the factors influencing usage intention toward mobile financial services. *Journal of Business Research, 65*(11), 1590-1599.

Yang, S., Lu, Y., Gupta, S., Cao, Y., & Zhang, R. (2012). Mobile payment services adoption

across time: An empirical study of the effects of behavioral beliefs, social influences, and personal traits. *Computers in Human Behavior, 28*(1), 129-142.

Zhang, X. (2017). Exploring the patterns and determinants of the global mobile divide. *Telematics and Informatics, 34*(1), 438-449.

Zhou, T., Lu, Y., & Wang, B. (2010). Integrating TTF and UTAUT to explain mobile banking user adoption. *Computers in Human Behavior, 26*(4), 760-767.

Editor's Note:

This paper was selected for inclusion in the journal as an CONISAR 2018 Distinguished Paper. The acceptance rate is typically 7% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2018.

Appendix.1: Constructs and Items

Construct	Item	Reference
Skill (SK)	I am very skilled at using mobile payment services via WeChat. I know how to use mobile payment services via WeChat to get what I want. I know more about using mobile payment services via WeChat than most users.	Novak et al. (2000) Koufaris (2002) Kaur et al. (2016)
Perceived Task-Tech Fit (PTTF)	In helping complete my mobile payment tasks, the functions of WeChat are enough. In helping complete my mobile payment tasks, the functions of WeChat are appropriate. In helping complete my mobile payment tasks, the functions of WeChat fully meet my payment needs.	Zhou et al. (2010) Lin and Huang (2008)
Perceived Usefulness (PU)	Using WeChat enhances my daily productivity in China. I find WeChat useful in my daily activities in China. Using WeChat enhances my effectiveness in daily activities in China.	Davis (1989)
Machine Interaction (MI)	Mobile payment services via WeChat works fast. WeChat processes payment fast. WeChat operates at high speed.	Novak et al. (2000) Koufaris (2002) Kaur et al. (2016)
User Satisfaction (USF)	WeChat satisfies my need to explore mobile payment services in China. WeChat satisfies my need to explore China. WeChat satisfies my need to cultivate my skills to use mobile payment services.	Wu et al. (2010)
User Stickiness (USN)	I would stay longer on WeChat than others when using mobile payment services. I would prolong my stay on WeChat when using mobile payment services. I would use mobile payment services via WeChat as often as I can. I would use mobile payment services every time I need to use WeChat.	Wu et al. (2010)

Security, Privacy, and Legislation Issues related to Commercial Drone Deliveries

Sandra A. Vannoy
vannoysa@appstate.edu

B. Dawn Medlin
medlinbd@appstate.edu

Department of Computer Information Systems & Supply Chain Management
Appalachian State University
Boone, North Carolina 28608

Abstract

This paper addresses commercial drone usage in relation to consumer delivery of products. As these unmanned flying vehicles become more prevalent in today's society, they are being embraced as a mechanism to address the consumer's ever-growing demand for quick delivery of products. However, this promising opportunity to address consumer demand has led to less positive issues such as privacy and security. Furthermore, as is the case with most emerging technologically supported initiatives, implementation of drone delivery has far outpaced current drone legislation. This paper seeks to provide an overview of the current status of drone delivery in the United States, security and privacy concerns, and current legislation surrounding commercial drone activity.

Keywords: Commercial Drones, Security, Privacy, Legislation

1. INTRODUCTION

Unmanned aerial vehicles (UAV) or unmanned aircraft systems (UAS) are known in mainstream societies as drones. Drones are essentially flying robots that can be controlled remotely or fly autonomously through embedded software and sensors that interface with global positioning systems (GPS). These unmanned flying robots have been classified based upon their size, intended use, flight range, speed, power system, among other categories (Hassanalian & Abdelkefi, 2017).

Drones are receiving increased attention around the globe as there has been recent emphasis upon unmanned flying vehicles for commercial use, particularly in the area of consumer product delivery. However, little is known about the impacts of the commercial application of drones on the world as we know it. Existing research suggests that while safety has been considered in various governmental rulings, little thought has been given to issues such as privacy and security (Dorr & Duquette, 2016). This paper is intended to provide insight into issues surrounding the commercial use of drones, with particular emphasis upon privacy, security, legislation, and regulation.

2. BACKGROUND

New technology developments have historically outpaced our understanding of their associated issues, such as privacy, security, and legislation, as their futures are unstable and unknown. Therefore, it is not surprising that there is little to be found in current academic or contemporary literature investigating these matters in the context of the drone.

McNeal (2012) suggested that the emergence of drones into the general public in the United States occurred due to the FAA Modernization and Reform Act of 2012, which loosened restrictions and provided greater airspace for drone flight. Also, in 2015 the FAA granted hundreds of new exemptions for companies to operate drones in the commercial segment including insurance, construction, and agriculture, but most of these exemptions (over 90%) were granted to small businesses having fewer than 10 employees (Joshi, 2017).

Placing drones within the congested nature of commercial airspace in the United States has proven quite complex, and therefore, the United States continues to lag behind much of the world in the use of drones for commercial purposes (Atwater, 2015). Much of the world has quickly outpaced the United States in terms of the commercial use of drones by dramatically loosening governmental restrictions, with Poland and South Africa being notably aggressive (Smith, 2016). Nonetheless, the promise of drone usage within the commercial realm is great, with the global market expected to surpass $120 billion worldwide by 2021 (Joshi, 2017)

To understand their use in general, it is helpful to look at the historical context of drones (originally known as unmanned aerial vehicles or UAVs). In 1918, the United States army used unmanned aerial vehicles, which were known as "flying bombs" that could hit a target up to 64 kilometers. The functions of drones were later enhanced by the military in their use as weapons carriers. The military in general has found that drones can be used to prevent casualties of war as they can provide accurate surveillance information and precise strike zones. Additionally, they have the ability to discern between intended and unintended targets (Rao, Goutham, Maione, 2016).

Several industries have expanded upon the military's use including agriculture, energy, deliveries, rapid response and emergency services, real estate, photography and others. According to Peter Diamandis (2018) in his article "Top 10 Reasons Drones Are Disruptive" between the years 1980 and 2010 there was major growth in the drone industry by both consumers and businesses in the four following areas:

1. GPS: In 1981, the first commercial GPS receiver weighed 50 pounds and cost over $100K. Today, GPS comes on a 0.3 gram chip for less than $5.

2. IMU: An Inertial Measurement Unit (IMU) measures a drone's velocity, orientation and accelerations. In the 1960s an IMU (think Apollo program) weighed over 50 lbs. and cost millions. Today it's a couple of chips for $1 on your phone.
3. Digital Cameras: In 1976, Kodak's first digital camera shot at 0.1 megapixels, weighed 3.75 pounds and cost over $10,000. Today's digital cameras are a billion-fold better (1000x resolution, 1000x smaller and 100x cheaper).
4. Computers & Wireless Communication (Wi-Fi, Bluetooth): No question here. Computers and wireless price-performance have gotten a billion times better between 1980 and today.

UAVs have been employed in a wide variety of civilian, law enforcement and military applications, the overwhelming majority of them beneficial. They can be used, for example, to deliver much needed medicines to remote areas, help rescuers identify people in need of assistance following a natural disaster, or to provide vital overhead imagery to police officers attempting to defuse a hostage standoff. In the commercial world, UAVs can be employed for a multitude of tasks as diverse as surveying, crop spraying, and traffic congestion monitoring. Scientific applications include air quality assessment, wildlife tracking, and measuring the internal dynamics of violent storms (Kwon, Kim, and Park, 2017).

UAVs also generate a number of economic benefits, with the potential to create thousands of jobs around the world that involve the design and production of UAVs as well as and spurring advances in robotics that will apply well beyond aviation, in fields ranging from manufacturing to surgery (Chamata, 2017).

3. COMMERCIAL DRONE USE

According to E-Marketer (2016), worldwide retail sales is approximately $26.6 trillion with online retail sales expected to grow to around $4 trillion by 2020. This means that e-commerce sales will account for almost 15% of the total retail market. The upward tick in e-commerce also means that fewer consumers are going into brick and mortar stores, and it is expected that they will receive their packages at the office or at their homes. Accordingly, there may be increasing opportunities for companies to use drones for the delivery of packages.

Major online research firm Skylark Services suggests that in any given day 110 million online orders are placed, with 100 million of the

products ordered weighing under five pounds. This evidence suggests that there is huge economic potential for drone delivery, and in fact, Skylark predicts a major disruption of the delivery world as we know it (Jenkins, Vasigh, Oster & Larson, 2017). Furthermore, drone delivery offers an interesting solution to the "last mile" problem faced by e-commerce companies wishing to reduce delivery times, reduce costs, and improve customer satisfaction (Murray & Chu, 2015).

In late 2013, Amazon announced its intention to implement a global drone-enabled delivery system, Prime Air. Since that time, a number of companies in the United States have expended a tremendous effort in developing a safe and reliable drone delivery system. Drone delivery company Flirtey has completed a number of FAA-approved drone deliveries, including medical supplies to the Remote Area Medical health clinic in Wise, Virginia in 2015, a delivery to a customer home in collaboration with 7-Eleven in 2016, and most recently has been engaged in delivery of pizzas with the Domino's Pizza company (Flirtey Continues to Lead Drone Delivery Industry, 2017). With each of these expansions of drone usage, additional benefits can occur that include the economic benefits such as the creation of new job titles and new innovations within the areas of robotics.

According to the Association for Unmanned Vehicle Systems International (AUVSI), in 2016, there were 150 drone manufacturers in the United States (Villasenor, 2016). With the expectation of more manufacturers to come, there is also the anticipation of lower costs and more product choices. Currently, there are increasing numbers of available UAV's under $1000, and it is expected that it is only a matter of a short time before GPS and video-equipped UAV's will be below $100. As with most technologies, there are advantages that may sound enticing, but there are several disadvantages which certainly include privacy and security.

With new technology use comes uncertainties and risks. Whether for personal or commercial use, issues surrounding privacy and security are paramount as unmanned vehicles can take photographs, videos, capture information and perform a number of other activities that can invade a person's privacy or not secure a person's information.

4. PRIVACY AND SECURITY ISSUES

The concept of privacy means separation from others and entails the ability to exclude oneself or exclude information about oneself. Furthermore, privacy as a concept fluctuates on national, individual, and cultural individualities (Serbua & Rotariua, 2015). Security, like privacy, has different meanings in different contexts. Arnold Wolfers (1952) states that the meaning of security is 'the absence of threats to acquired values' which appears to capture the basic intuitive notion underlying most uses of the term security, and can be applied to many different generic situations. Privacy and security as related to drone technology are further complicated by its nascent nature, with few clear rules or regulations.

When drones are used commercially in a society, a number of public safety and privacy issues must be addressed (Jansen, 2013). With new technologies, comes uncertainties and risks, and issues such as security, privacy, and legislation take a back seat to the innovation of the technology itself. While unmanned vehicles can take photographs, videos, capture information and perform a number of other activities that can invade a person's privacy or not secure a person's information, we currently have little understanding of these issues or how to address them.

Drone activities have given rise to activities and possible threats to data privacy. One of the largest privacy concerns is that drones can be used for spying given the fact that most drones have cameras attached to them or embedded within the technology of the drone. Capturing data about a property or peeking into someone's home is an easy task for drones. In addition, drones are capable of capturing large amounts of data that can easily include pictures and videos. Other drone activities can include the capturing and exploitation of data in conjunction with other data and the use of facial recognition to identify individuals.

Overt or unwanted surveillance by drones can impinge upon an individual's right to physical privacy. The privacy of personal behavior is about the freedom of an individual to behave as they wish without worry or concern of undue observation and interference from others (Clarke, 2014). A surveillance target may be an area, or one or more objects, as mentioned earlier, including people (Wigan and Clarke, 2006). A combination of physical surveillance with data gathering can provide for the acquisition of

information about a person's location at a certain time or place which constitutes tracking and therefore means that inferences may can be drawn about patterns of the individual's behavior. These inferences can then be used for criminal activity such as, stalking, blackmail and harassment.

Drones can be targeted for software and hardware attacks because they fly, capture video, and can be remotely controlled. They can be also targeted for command and control data link jamming and spoofing, in which a hacker can block or falsify the data link in order to disrupt or take control of the device. Navigational sensor jamming can also disrupt and take over navigation. Additionally, hackers can tap the video or photo link, where they intercept the video and other data from the drone.

Drones can also crash, thus allowing others to not only confiscate a product, but also gain access to the receiver's name, address and phone number as well as goods and other information within the package. Draper (2015) states that crashes are inevitable because of issues related to weather, other aerial vehicles, buildings, and birds, or hackers who may gain control of the drone, leading to safety concerns for people and property.

Drone units are vulnerable to two different kinds of security attacks that can occur on their GPS navigational systems. 'Spoofing' entails the sending of strong (but fake) GPS signals towards a drone, so that it is essentially "hijacked" instead of following its programmed directions. The drone can then be manipulated to crash or be flown to another location such as the attacker's or another specified location. This could make it possible for an employee at Amazon to be held responsible for the consequences of the "spoofed" drone since it is very difficult to prove the origin of the navigation signals. It wasn't until 2014 that a successful spoofing attack was conducted against a drone by a researcher at the Department of Homeland Security facility. For now, not all commercial drones use encryption methods that render them invulnerable to any currently known spoofing attack, but even those enabled with encryption are still left susceptible to 'jamming.' In a jamming attack, the drone is overwhelmed with signals to the GPS antenna. The encryption ensures that no fake signal is mistaken for the true one, but the true signal also cannot get through. Unintended collisions seem to be unavoidable in such scenarios, especially in an unregulated environment (Rao, B., Gopi, A., & Maione, R., 2016).

The advantages of drone delivery may sound enticing, but there are several disadvantages concerning privacy and security issues. The U.S. Federal Trade Commission (FTC) has raised several questions surrounding the topic of privacy and security, as FTC researchers were able to hack into three different off-the-shelf drones. Furthermore, they were able to take control of the camera feed on each drone; for two of the drones, they were able to turn off the aircraft to make it fall from the sky and seize complete control of the flight path (Glass, 2016).

Just like military operations, commercial companies such as Amazon can use the data collected from drone deliveries and returns in order to assist in their marketing campaigns. According to Jeff McCandless, Founder and CEO of project44, "Amazon can leverage information about your vehicles, the exterior of your home and any property visible from the outside, and use that to market-related products to people. They can even obtain information about when people are home, when they are outside, etc. There's no telling what other ideas they'll come up with as they bring in rounds of data and begin analyzing it. That said, one has to wonder where it ends." This, of course, raises issues about the use of consumer data without their knowledge or consent.

Under President Obama's administration, Congress held hearings related to issues around the use of drones, with over half of the states enacting some type of drone legislation after the fact. But once again, the issues of privacy and security were not directly addressed. In fact, in every state where laws were passed, the new legislation focused more on the technology itself, rather than the harm that surveillance could create (Thompson, R. 2015).

Justice Samuel Alito wrote a concurrence in United States v. Jones, in a January 2012 Supreme Court ruling that addressed the constitutionality of affixing a GPS tracking device to a vehicle without a valid warrant: "[i]n the pre-computer age, the greatest protections of privacy were neither constitutional nor statutory, but practical." Although Justice Alito's statement was directed toward GPS tracking alone, it has much relevance to potential invasions of privacy through use of UAVs. In comparison with manned aircraft, UAVs can be very inexpensive to procure and operate. As the practical barriers such as cost to obtaining aerial imagery are reduced, the likelihood of their use becomes stronger. Consequently, concerns about the ways they are used become heighted as well.

Drones lead to a range of issues not seen with many emerging technologies. Primarily, there are few clear rules or regulations holding manufacturers responsible for incorporating security measures that might prevent tampering by malicious hackers (Glaser, 2016). Currently, organizations are more concerned with how this phenomenon can be incorporated into business models than with issues of privacy and security.

5. LEGISLATION AND REGULATION

A variety of laws may be applicable to drones and their usage including trespassing, publication of public facts, and stalking and harassment (Vallesenor, 2013). A complicating factor is that different localities may have differing laws related to airspace usage according to federal legislation. The FAA enacted the FAA Modernization and Reform Act of 2012 (FMRA) that called for the integration of unmanned aircraft systems (UAS), or "drones," into the national airspace by September 2015. Unfortunately and during that time, "the substantive legal privacy framework relating to UAS on the federal level has remained relatively static: Congress has enacted no law explicitly regulating the potential privacy impacts of drone flights, the courts have had no occasion to rule on the constitutionality of drone surveillance, and the Federal Aviation Administration (FAA) did not include privacy provisions in its proposed rule on small UAS" (Thompson, R., 2015). Under federal law, all UAVs must apply to the FAA for permission to fly unless they fall under the exception clause. The process for obtaining permission to operate drones differs depending on whether the operator is a public operator or a private commercial operator.

One of the key takeaways from the 2012 legislation is the visual line-of-sight (VLOS) mandate. VLOS ensures the pilot will only operate the drone as far as he or she can see. While everyone's vision is different, this ensures that the drone wouldn't legally be able to travel very far or encounter obstacles unknown to the operator. To realistically use drones in commercial deliveries, it will be quite difficult if not impossible to always maintain a line of sight. Therefore, it is assumed that newly adopted FAA regulations may relax some of the regulations for specific classes of UAS operations (Schlag, 2017).

In December 2015, the FAA passed a federal law requiring all drones weighing over 250 grams, or a little over one-half pound, and their users to be registered online. The law has been justified due to privacy and public safety concerns, as the FAA had reported 1133 cases of unsafe use (FAA.gov). Due to the increasing number of UAVs it was posited that with this increase comes the possibility of technical failure either due to the failure of the technology or users' inexperience. As a result of this law, a user flying / owning a without a certificate, and even on their own property, can face both civil and criminal sanctions including fines and imprisonment.

Most recently in October 2017, President Donald Trump signed a memo to the Department of Transportation (DOT), directing them to begin the process of developing rules to allow commercial drone operators to fly more freely in the U.S. The memo directs the DOT to take proposals from local, state, and tribal leaders over several months, and then select the five most promising proposals and run small experiments over the next three years, to see which one is the best solution. Then, the winning proposal will be implemented nationally (Stewart, 2017).

Though the FAA may not have strict rules for drone use pertaining specifically to privacy issues, many states and localities have strict *Peeping Tom* regulations that may apply if a drone were to hover over private residences. Again, the FAA is relying on local law enforcement agencies to address such issues at this time as they arise. In a study conducted by the Center for the Study of the Drone at Bard College, they found that at least "…130 localities in the U.S. have their own local drone rules, which have extended beyond the rules implemented by the FAA" (Bard College Surveys Legal Cases Involving Drones, 2017).

Drone delivery presents a particular challenge for law enforcement officials as drone use has increased dramatically, making it difficult for the FAA to monitor their flights as it does with commercial airlines. Unlike manned airlines, drones can be operated almost anywhere and are not supervised by traffic controllers. Additionally, Senator Ed Markey, of Massachusetts introduced in March of 2017, legislation to put in place more privacy protections against drones and their invasions upon individuals and their privacy. During the same period, "The Drone Aircraft Privacy and Transparency Act" was also introduced in the House with the support of Rep. Peter Welch of Vermont.

Outside of the United States' legal system is an international framework, the International Covenant on Civil and Political Rights (ICCPR). In some countries, civil rights may be protected by their constitution; however, some of these rights are insufficient to significantly curb the use of

drones in the area of visual surveillance. In the United States, the Fourth Amendment is primary to the issue of privacy and UAS operations. Under the Fourth Amendment, Americans are guaranteed a certain "right" to privacy through the right "to be secure in their persons, houses, papers, and effect against unreasonable searches and seizures" (U. S. Const. amend. IV). There are dissenting opinions, however, concerning the strength of the Fourth Amendment in relation to consumers and their privacy protections from the use of drones. Some advocates of the U.S. Constitution believe that there will be a much stronger measure of protection against government UAS privacy abuses than is widely appreciated, while others suggest that that there is further need for substantial statutory and common law protections that will protect individuals and their privacy rights.

According to some legal scholars, drones, with their current and projected capabilities, present a perfect storm of issues that fall outside of the current Fourth Amendment jurisprudence, but still appear relevant to the Fourth Amendment (Bomboy, 2014). As drones can travel on public airways at low or high altitudes, undetected and with little or no undue noise, and drones can use technologies to gather an abundance of intimate details and information, it has been suggested that law enforcement will likely increasingly use drones for domestic surveillance, and all of these actions will likely propel drones to the forefront of courts' dockets.

In February 2018, a helicopter crash-landed in South Carolina, with the crash being triggered by a civilian drone. Neither the drone nor owner of the drone were ever found. As drones become more popular, incidents such as this one will most likely be on the rise. Though it is noted to be the first drone-related crash of an aircraft in the U.S., it is expected that more of these occurrences will happen as more and more drones are being purchased for both recreational and entertainment use (Bloomberg, 2018). Though this may have been the first noted crash, there have been a number of other incidents that have created serious and almost deadly results. As examples, a commercial jet and a drone came within 200 feet of colliding near Los Angeles' LAX airport in March 2016 and a JetBlue pilot taking off at JFK Airport reported a near collision with a drone at about 5,800 feet in January of 2017. The FAA chronicled 583 near misses between aircraft and drones between Aug. 21, 2015, and Jan. 31, 2016. That averages out to approximately 116 reported incidents monthly, with that number increasing (FAA.gov, 2017). The US Department of Transportation estimates that by the year 2035 175,000 unmanned aircraft will be used for commercial purposes, surpassing the number of manned aircraft (Volpe, 2013). This statistic emphasizes the impending disruption to traditional commercial airspace and the need for major legislative and regulatory attention with regard to commercial used of drones.

In order to assist consumers with information so that they do not break the law, the Federal Aviation Association is currently leading an outreach campaign to make end-users aware of the privacy and safety issues surrounding drone usage (U.S. Department of Transportation, 2018). The website offers guidance for anyone operating a UAS by providing information about airspace restrictions, and how not to endanger individuals or other aircraft. Additionally, the site offers No Drone Fly campaign materials, and an application called B4UFLY Mobile App that assists in determining airspace restrictions and other flying requirements based on an end-user's GPS location.

6. CONCLUSIONS

Little research has been done to understand the impact of the drone in commercial activity. The rapid growth of drone use by both civilians and businesses has created a number of challenges that include privacy and security concerns as well as legislative and regulatory issues. Existing regulations do not fully address the impact of these unmanned flying objects, and there is also the strong potential for intentional as well as unintentional misuse. Drones will inevitably be a key part of our trillion-sensor future, transporting a variety of sensors (thermal imaging, pressure, audio, radiation, chemical, biologics, and imaging) and are already connected to the Internet. This opens up possibilities for device-to-device communication, with operators, as well as introduces the potential for hacking.

These authors intend to extend this paper by incorporating a study of user acceptance of the drone as a "last mile" delivery mechanism. Similar to the rationale of Im, Kim, & Han (2008), we hypothesize that perceived risk and technology type will be important factors in users' acceptance of the drone as a delivery to the home mechanism.

Additional future research into the impact of the drone is needed. For example, an economic impact analysis of the impact of drone delivery upon traditional commercial delivery could help prepare the package delivery industry for the

future. Additionally, economic impact analysis could help identify new revenue streams and job opportunities. A host of other issues could be examined to lead to a better understanding of commercial drone usage, including the social, political, and cultural contexts.

7. REFERENCES

Bamburry, D. (2015). Drones: Designed for product delivery. Wiley Online Library. Retrieved on May 19, 2018 from http://onlinelibrary.wiley.com/doi/10.1111/drev.10313/pdf.

Bard College Surveys Legal Cases Involving Drones. (2017). Retrieved on on June 5, 2018 from http://www.aero-news.net/index.cfm?do=main.textpost&id=74ea7944-ff2b-4e96-b5f8-4e9e09958f17.

Bloomberg News. (2018). This might be the first Drone-Related Aircraft Crash. Retrieved on February 25, 2018 from http://fortune.com/2018/02/16/south-carolina-drone-helicopter-crash/.

Bomboy, S. (2014). A Legal Victory for Drones Warrants a Fourth Amendment Discussion, NAT'L CONST. CTR. Retrieved on February 9, 2018 from http://blog.constitutioncenter.org/2014/02/acourt-victory-for-drones-warrants-a-fourth-amendment-discussion.

Chamata, J. E. (2017). Convergence of the Unmanned Aerial Industry. Theoretical Economics Letters, 7(02), 175.

Clarke, R. (2014). The regulation of civilian drone's impacts on behavior privacy. Computer Law & Security Review, 30(1), pp. 286-305.

FAA.gov (2017). FAA Releases Updated Drone Sighting Reports. Retrieved on February 27, 2018 from https://www.faa.gov/news/updates/?newsId=87565on.

Im, I., Kim, Y., & Han, H. J. (2008). The effects of perceived risk and technology type on users' acceptance of technologies. Information & Management, 45(1), 1-9.

Jansen, B. (2013) A Look at the Privacy Policies for FAA's Six Drones Test Sites. ACLU. Retrieved on February 17 from https://www.usatoday.com/story/news/nation/2013/12/30/drone-test-sites/4248771/.

Kwon, H., Kim, J., & Park, Y. (2017). Applying LSA text mining technique in envisioning social impacts of emerging technologies: The case of drone technology. Technovation, 60, 15-28.

Laguna, J., & Marklund, M., Business Process Modeling, Simulation, and Design, Prentice Hall, New Jersey, 2005.

McNeal, G. (2012, April). A primer on domestic drones: Legal, policy, and privacy implications. Forbes. Retrieved from www.forbes.com/sites/gregorymcneal/2012/04/10/a-primeron-domestic-drones-andprivacy-implications/.

Michel, H.A. & Gettinger, D. (2017) Drone Incidents: A Survey of Legal Cases. Retrieved from http://dronecenter.bard.edu/files/2017/04/CSD-Drone-Incidents.pdf on February 9, 2018.

Pogue, David. (2016). Amazon reveals details about its crazy drone delivery program. Yahoo Tech. Retrieved on March 5, 2018 from https://www.yahoo.com/tech/exclusive-amazon-reveals-detailsabout-1343951725436982.html.

Slove, D. (2006). A taxonomy of privacy. University of Pennsylvania Law Review, 154(3), pp. 477-56-.

Smith, G. (2016, May). Here Comes the Latest Drone Army. Retrieved from http://fortune.com/2016/05/09/here-comes-the-latest-drone-army/.

Stewart, J. 2017. PRESIDENT TRUMP MOVES TO FILL AMERICA'S SKIES WITH DRONES. Retrieved on February 17, 2018 from https://www.wired.com/story/faa-trump-drones-regulations/.

U. S. Const. amend. IV.

United States v. Jones, 132 S. Ct. 945, 963 (2012) (Alito, J., concurring in the judgment).

United States Department of Transportation (2018). Where to fly. Retrieved from https://www.faa.gov/uas/where_to_fly/ on February 17, 2018.

Vallesenor, J. (2013). Observations from above: unmanned aircraft systems and privacy. Harvard Journal of Law Public Policy, 36(2), 457-517.

Wigan, M., & Clarke, R. (2006). Social impacts of transport surveillance. Prometheus, 24(4), 389-4

Effects of Normalization Techniques on Logistic Regression in Data Science

Adekunle Adeyemo

Hayden Wimmer
Hayden.wimmer@gmail.com
Georgia Southern University
Statesboro, GA 30458

Loreen Powell
lpowell@bloomu.edu
Bloomsburg University
Bloomsburg PA 17815

Abstract

The improvements in the data science profession have allowed the introduction of several mathematical ideas to social patterns of data. This research seeks to investigate how different normalization techniques can affect the performance of logistic regression. The original dataset was modeled using the SQL Server Analysis Services (SSAS) Logistic Regression model. This became the baseline model for the research. The normalization methods used to transform the original dataset were described. Next, different logistic models were built based on the three normalization techniques discussed. This work found that, in terms of accuracy, decimal scaling marginally outperformed min-max and z-score scaling. But when Lift was used to evaluate the performances of the models built, decimal scaling and z-score slightly performed better than min-max method. Future work is recommended to test the regression model on other datasets specifically those whose dependent variable are a 2-category problem or those with varying magnitude independent attributes.

Keywords: Normalization, Logistic Regression, Z-Score, Min-Max, Decimal Scaling

1. Introduction

Advancements in the field of data science have allowed the application of several mathematical concepts to behavioral patterns of data. Precisely, different normalization techniques have been applied to numerous datasets to solve problems from all walks of life. Data normalization is a preprocessing method used in different data mining systems, particularly, for classifying algorithms such as neural networks, clustering and neighbor classification (Evans, 2016). A lot of works have been published in data normalization and its application to different fields of human endeavors; Statistical Normalization and back Propagation for Classification, Min-Max Normalization based on Data Perturbation method for Privacy Protection, Importance of Data Normalization for the application of Neural Networks to Complex Industrial Problems and the Impact of Normalization Methods on RNA-Seq Data Analysis. In this research, we investigated how different normalization techniques affect the Performance of a Logistic Regression Classifier. Logistic regression is an ideal tool for answering classification questions. It is a model that can be used to forecast the binomial outcome of a dependent (target) variable using one or more independent (predictor) variables. The independent variables can be binomial, numerical or even categorical. Logistic Regression algorithm is used to classify Red Wine dataset based on its

quality, the dataset was then normalized using three different normalization methods and different models were built as a result. These new models were compared with the baseline model and performance effect was then discussed.

The original dataset is modeled using SQL Server Analysis Services (SSAS) Logistic regression tool. This serves as the baseline model, this is followed by describing the normalization methods used in transforming the data and different logistic regression model are then built as a result.
Since the aim of this study is to compare how various normalization techniques affect the performance of logistic regression model, the most commonly used normalization methods; Min-Max, Z-score, and Decimal Scaling are used to transform the original data and the performance of the resulting models are evaluated using the accuracies and model lifts as the major metrics. The remaining format of this paper is the following: literature review, methodology, results, implications, and conclusion.

2. Literature Review

There have been different publications on data normalizations and how different normalization methods are applied in different fields to solve various problems. The publications in this category are described in the subsequent paragraphs.

Min-Max normalization techniques was used to preserve privacy of data as a distorting method by (Jain & Bhandare, 2011). Min-Max normalization technique was applied to the original dataset (M) to get a newly transformed dataset ($\bar{M}$) with same number of rows (records) and columns (attributes). The $\bar{M}$ can appear as a distorted form of M. $\bar{M}$ was then altered further to improve its security by multiplying it with a negative number. This action changed the other and values of $\bar{M}$ as positive numbers become negative. This technique was applied to four different real-life databases obtained from UCI Machine Learning Repository. This data perturbation method with shifting factor SF = -15 was applied on these real-life databases. The experiments conducted showed that the value difference (VD) and accuracy of two of the datasets changed with respect to SF. Another SF was carefully selected to better the result. The publication is important to this research as it described how data normalization technique was being used to change the meaning of dataset to preserve its' privacy.

Normalization methods as they relate to sequencing of RNA data and Impact analysis of the results of gene expression were compared by (Zyprych-Walczak et al., 2015). Five Normalization methods were compared using three real-life RNA-seq datasets. Housekeeping Genes (HG) was selected as the analytical criterion for comparing the normalization methods used in processing of RNA-seq data. Since the goal of the study was to find out how normalization techniques impact differential expression results, differential analysis was conducted using edgeR method in the edgeR Biconductor Package after the application of each normalization approach. The results of the experiments conducted were compared using different factors. These results showed that the impact of the normalization technique depends on the data structure and the criteria for comparison. This study opens explores the fact that the influence of data normalization method is dependent on the dataset and the criteria for comparing the performance.

How input data normalization improve the performance of parameter predictors trained to assess the value of several attributes of a nuclear plant was showed by (Sola & Sevilla, 1997). Two different systems were studied, pressurizer pressure and power transferred between the reactor coolant system and the main vapor system. These two networks were studied using neural network simulator SINAPSIS. Three-layered perceptron was used in both systems and training was done through back propagation algorithm. 6 and 8 input variables were used accordingly. The influence of network architecture on the results was studied by evaluating the behavior of a wide range of options. The input variables were normalized using five different normalization techniques. The results showed that a suitable normalization of input variables before network training reduced estimation errors by 10% and the required calculation time during the training process is also reduced. This study proves that normalization of input data can improve the performance of neural network classifier.

Different normalization methods applicable in back propagation neural networks as they enhance the reliability of the trained network was presented by (Jayalakshmi & Santhakumaran, 2011). The reliability of each of the method described was stated and how they affect the attributes of the datasets. The simulations were conducted using MATLAB, different networks were reproduced and experiment with. The network was trained 10 times and the

performance was examined at different periods. The results of the experiments clearly revealed that the performance of the dataset used in the classification model relied on the normalization methods and that the Statistical Column normalization produced the most accurate result. The study showed how the performance of back propagation neural nets can be improved upon using some applicable normalization approaches.

Other relevant literatures to the work at hand described data normalization protocols and features specifically constructed to improve the performance of some classifiers, which are important to our research topic. Specifically, a protocol for data exploration that can be used to avoid common statistical problems was proposed by (Zuur, Ieno, & Elphick, 2010). The protocol in question was divided into eight linear flexible stages. The stages include identifying and removing outliers, variance homogeneity, normal distribution of data, lots of zero in the data, the existence of correlation between covariates, considering the relationships between response and predictor variables, considering interactions between output attribute and different type of covariates and independent observation of response variable. The paper discussed a series of drawbacks that can impact the output of an analysis, but these can be avoided using the systematic data exploration procedure described before undertaking any analysis. This paper is such an important one as it teaches how to prepare dataset before applying it in analysis or modeling.

Kaizen Programming (KP) approach was employed to improve Logistic Regression model to find high-quality nonlinear combinations of the original features in a dataset by (de Melo & Banzhaf, 2016). KP together with LR model was used to filter important features of credit scoring dataset and Akaike Information Criterion (AIC) was used as selection model aimed at improving the prediction performance of LR. The performance of KP was implemented using Australian Credit Approval dataset, the continuous variables in the dataset were discretized since the implementation did not work with mixed type attributes. Before models were built, identical or highly connected features were discarded. KP was implemented in Python and experimental analysis was executed on Weka using the LR as the classifier. The new dataset with the best accuracy for each desired feature was selected. The experiment showed that KP results were competitive though some imbalanced because it generated different features compared to other methods in the literature. The study proved that Logistic Regression model can be used together with another problem-solving method such as KP to improve its predictive performance.

A recurrent neural network approach was applied to stock price pattern recognition by (Kamijo & Tanigawa, 1990) . The proposed network was a four-layered architecture with one layer for input, two as hidden and one layer as output. The output layer is used to discriminate nonlinear patterns. Sixteen experiments were conducted with sixteen stock price patterns for recognition. After the experiments, the actual pattern was correctly recognized 15 times out of 16 experiments that were conducted. The results of the experiments showed that normalization by exponential smoothing introduced s bias which is the difference in name and time span. The research work showed that exponential smoothing way of normalizing data introduced some errors to the neural network model for pattern recognition.

3. Methodology

This research work investigates the effect of different Normalization Techniques on the prediction accuracy of Logistic Regression model. The SQL Server Analysis Services (SSAS) is the major tool used for the work. SSAS is a tool from the Microsoft Business intelligence team, for developing Online Analytical Processing (OLAP) solutions. A typical workflow consists of authoring a multidimensional or data model in tabular format, deploying the model as a database to an SSAS or Azure Analysis Services. SSAS environment is a collection of machine learning algorithms such as, Neural Network, Decision Tree, Naïve Bayes, Logistic Regression and so on.

Dataset

The problem that was selected for this research is to predict the quality of red wine data using Logistic Regression model of SQL Server Analysis Services (SSAS). To investigate the performance of this classifier, the model was applied to Red Wine Quality dataset of the Portuguese "Vinho Verde" wine collected from UCI Machine Learning Repository. The output attribute is a 11-class problem between 0(very bad) and 10(very excellent) for red wine quality. The dataset consists of 1599 instances. Each record consists of 11 input attributes. The relevant independent attributes determined by the dependency capacity of Logistic Regression Model are:

1. Alcohol
2. Sulphates
3. Fixed Acidity
4. Citric Acid

5. pH Values (objective test)
6. Residual Sugar
7. Free Sulfur Dioxide

There are 10 instances of the dataset with quality of 3, 53 with quality of 4, 681 with quality of 5, 638 with quality of 6 and 18 with quality of.

Structural Explanation of a Logistic Regression Model

The SSAS Logistic Regression model is formed using Neural Network algorithm with the elimination of hidden node. Hence, the general model for a logistic regression is almost the same as that of neural network; each model has a single root node representing the model and the details about it, and a distinct marginal statistics that gives the details about the independent attributes used in the model.

Furthermore, the model consists of a subnetworks for each dependent attribute. Each subnetwork contains two branches; one for the input layer and the other contains the hidden layer and the output layer. However, in this model, the hidden layer is empty as it has no children. So, the model consists of nodes that stand for individual outputs and inputs with empty hidden nodes. As it is shown in Figure 1, the logistic regression model is presented using the Neural Network Viewer. The neural network viewer allows the filtering of input attributes and their values and graphical view as these affect the outputs. There are various tabs in the viewer that show the probability and lift association as regards to the input and output values.

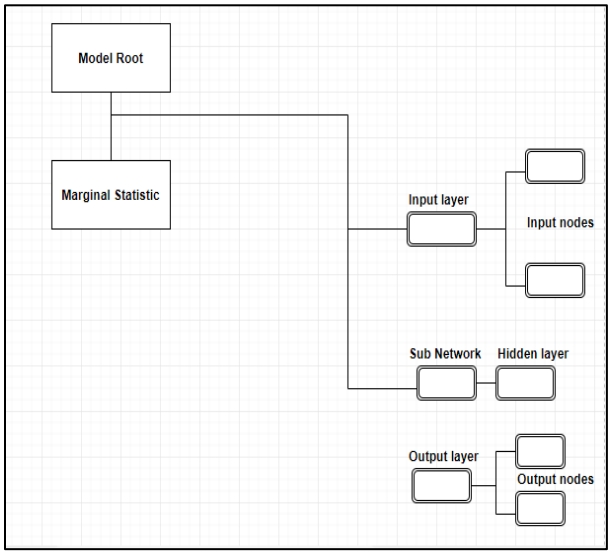


Figure 1: Logistic Regression Model

Normalization Techniques

The three types of Normalization techniques applied to the dataset are described in the subsequent paragraphs. There are numerous types of Normalization methods but these three are chosen based on their popularity in the reviewed literatures. These techniques include: Min-Max, Z-score and Decimal Scaling.

Min-Max Normalization

This technique is a strategy that linearly transform the attributes or outputs from one range of values to a new range of values. Mostly, the variables are transformed to lie between 0 and 1 or -1 and 1. The rescaling is usually achieved using the linear transformation given as:

$$y = (x - \min(x))/(\max(x) - \min(x))$$

Where min and max are the minimum and maximum values in X, where X is the set of observed values of x. In other words, max(x) – min(x), is the range of X. The advantage of this normalization method is derived from the fact that all relationships in the data are exactly preserved.

Z-Score Normalization

This method is the most popular normalization method which converts all input values to a common measure with an average of zero and standard deviation of one. The mean and standard deviation are calculated for each attribute. Each value of an attribute X is normalized using the computed mean and standard deviation. The transformation equation is given as:

$$y = (x - \text{mean}(X))/\text{std}(X)$$

Where mean(X) = mean of attribute X and std(X) = standard deviation of attribute X. The advantage of this method is deduced from the fact that it reduces outliers' effect on the data.

Decimal Scaling

This normalization technique works by moving the decimal point of values of attribute X. The number of points moved is determined by the maximum absolute value of X. The value x of attribute X is normalized to y by using the formula:

$y = x/10i$

Where i is the smallest integer that satisfy the condition $\text{Max}(|y|) < 1$.

Structural Representation of Activities

Figure 2 shows the structural representation of all the activities involved in this research. The dataset is presented into Logistic Regression (LR) model in four different views and structures. The

original dataset and the resulting datasets after normalization using three different techniques as depicted in the model above served as inputs to the LR model. Four different prediction models were generated as outputs. The three outputs from the normalization techniques were compared with the output of the original dataset based on the prediction accuracies of the model.

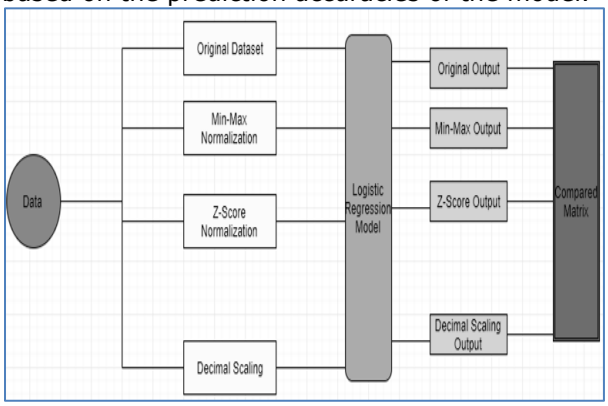


Figure 2: Structural Activities Model

4. Results

The tests to evaluate how different normalization techniques affect the performance of Logistic Regression Model (LRM) have been conducted. The Red Wine Quality (RWQ) dataset which consists of 1599 records was divided into two sets using SSAS LRM. The training set consists of 70% of the original dataset (1119 records) and the testing set consists of the remaining 30% of the dataset (480 records). The RWQ dataset was normalized using the Min-Max, Z-Score and Decimal Scaling normalization methods as described under the method section. These same percentages of training and testing sets were used for each of the normalization techniques.

		Actual					
		6	3	8	7	4	5
Predicted	6	92	2	4	41	7	41
	3	0	0	0	0	0	0
	8	0	0	0	0	0	0
	7	9	0	2	14	1	3
	4	0	0	0	0	0	0
	5	68	2	0	11	10	172

Table 1: Classification Matrix for Baseline-Model on Quality

Tables 1, 2, 3 and 4 show the classification matrices for the four models, that is, one for the original dataset and the remaining three for each normalization techniques. These matrices are also called Confusion Matrices and are used for summarizing the performance of a classification algorithm or classifier. The columns of the classification matrices correspond to actual values, rows correspond to predicted values.

Accuracy and lift were the two major metrics used for evaluating the performances of the LR models.

		Actual					
		6	3	8	7	4	5
Predicted	6	106	1	2	40	3	54
	3	0	0	0	0	0	0
	8	0	0	0	0	0	0
	7	11	0	4	14	1	4
	4	0	0	0	0	0	0
	5	77	2	0	2	7	151

Table 2: Classification Matrix for Min-Max on Quality

		Actual					
		6	3	8	7	4	5
Predicted	6	112	1	3	33	8	53
	3	0	0	0	0	0	0
	8	0	0	0	0	0	0
	7	9	0	1	11	1	2
	4	0	0	0	0	0	0
	5	79	2	0	7	9	148

Table 3: Classification Matrix for Z-Score-Model on Quality

		Actual					
		6	3	8	7	4	5
Predicted	6	113	1	3	37	7	43
	3	0	0	0	0	0	0
	8	0	0	0	0	0	0
	7	13	0	3	12	1	3
	4	0	0	0	0	0	0
	5	76	2	0	8	9	148

Table 4: Classification Matrix for Decimal-Scaling on Quality

Accuracy

The accuracy of a model is defined as the percentage of the test dataset correctly specified. This is given as:

]

$$\text{Accuracy} = \frac{No\ of\ correctly\ classified\ test\ samples}{\text{Total no of Test Samples}}$$

Number of correctly classified test samples is the summation of all the diagonal values in a matrix. Table 5 contains the accuracies as reported by the four models. Figure 3 shows the graphical representation of the models accuracies.

Model	Accuracy
Data	57.92
Min-Max	56.46
Z-Score	56.46
Decimal Scaling	56.88

Table 5: Model Accuracy in Percentages

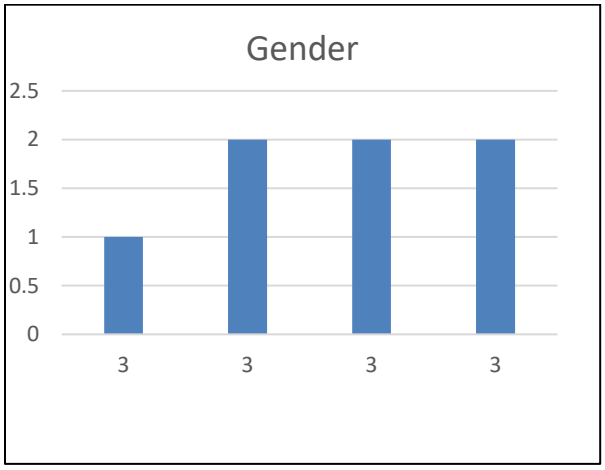


Figure 3: Graphical representation of the models accuracies

Model Lift

A lift measures the proportion of the true positives from the model compared to proportion of positive hits in the dataset overall. The lift of a model can be obtained directly from the SSAS by clicking on the Lift Chart tab under the Mining Accuracy Chart.

From table 5, the LRM performs better on the original dataset as the prediction accuracy was about 58% compared to when the dataset was normalized with accuracies of about 56%, 56% and 57% for Min-Max, Z-Score and Decimal Scaling normalization methods respectively. The LRM behaves similarly even when the training set was increased to 80% for all the normalization techniques.

The figures 4, 5, 6, 7, 8, 9, 10 and 11 showed the Lift Charts and Legends for our Baseline, Min-Max, Z-Score and Decimal-Scaling models respectively. In the figures, the legends show that two lines should be displayed, one for the specific model, such as, baseline model, and one for the ideal model.

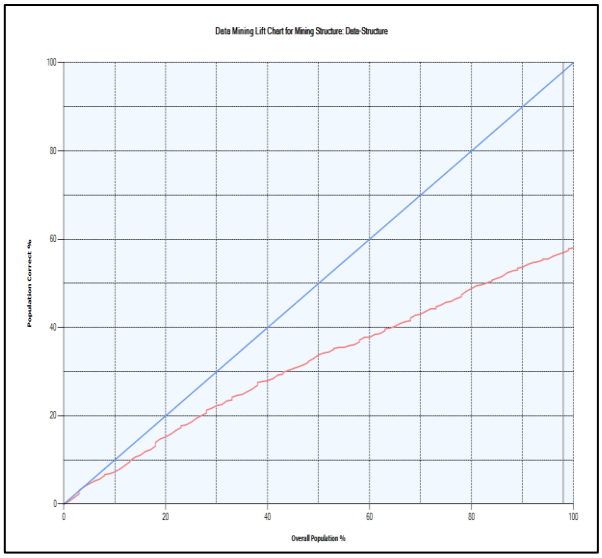


Figure 4: Baseline Model Lift Chart

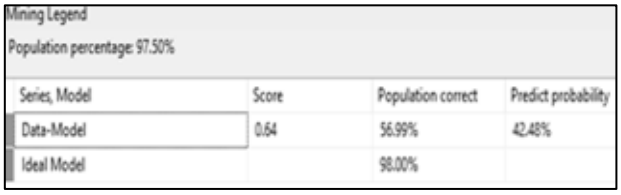


Figure 5: Baseline Model Lift Legend

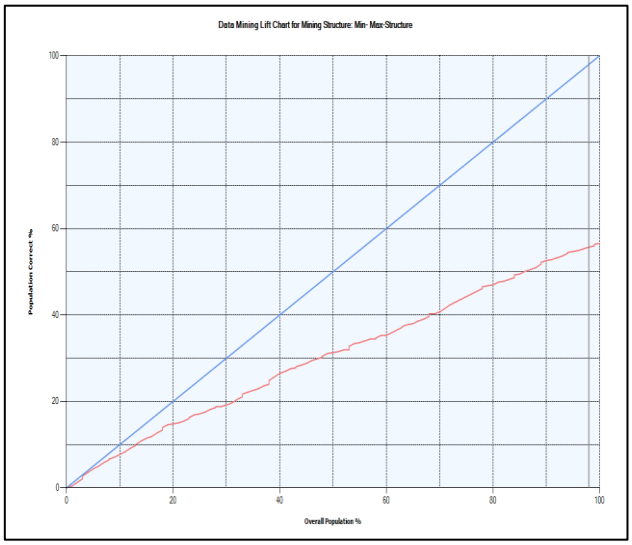


Figure 6: Min-Max Model Lift Chart

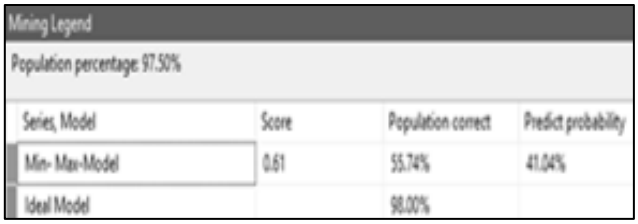


Figure 7: Min-Max Model Lift Legend

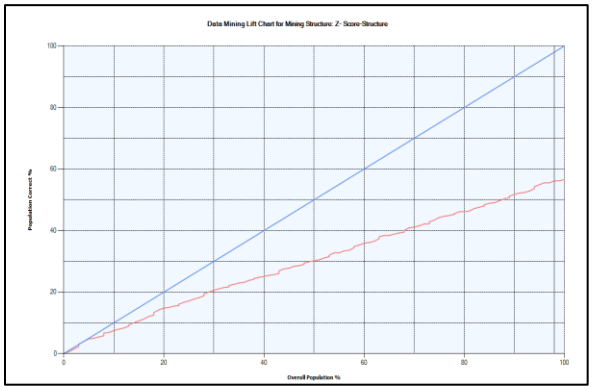


Figure 8: Z-Score Model Lift Chart

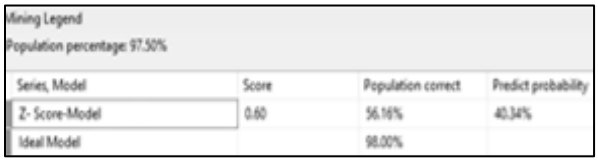


Figure 9: Z-Score Model Lift Legend

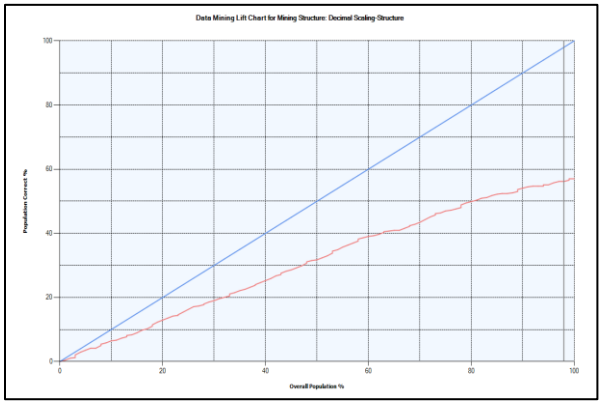


Figure 10: Decimal-Scaling Model Lift Chart

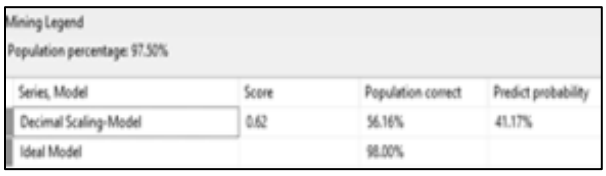


Figure 11: Decimal-Scaling Model Lift Ledged

For a perfect classification model, as we have in figures 4, 6, 8, and 10 the Ideal Model is on top of the specific model chart line. The solid gray vertical bar can be clicked and dragged horizontally to examine different values along the plotted line in the Legend windows. In our case, the Legend windows show that for 98 percent of the overall population 56.99%, 55.74%, 56.16% and 56.16% for Baseline, Min-Max, Z-Score and Decimal Scaling models respectively were correctly predicted.

Although the accuracy of a normalized dataset is expected to improve with classifiers such as Neural Networks, LRM performs poorly to normalized dataset. This LRM's behavior might be because of the magnitude of the independent variables in the dataset that are already close to one another before normalization or the output attribute that is an 11-class problem.

5. Implications for Practice

Data Normalization means transformation of all attributes in the dataset to a specific scale. We do data normalization when seeking for relationship between the variables in the dataset. Several works have been published on data normalization and how important these techniques have become as a data preprocessing strategy but little effort has been geared towards how these methods affect the performance of machine learning algorithms especially the Logistic regression. This research work is very important in that it will serve as the foundation for researchers to build on and for data scientists to see that Logistic regression performs poorly under the influence of normalized data.

6. Conclusions and Future Directions

The performance of Logistic Regression Model was evaluated with respects to three different normalization techniques. As usual, normalization of dataset is expected to improve the predictive accuracy of a machine learning algorithm but LR behaves poorly to the three normalization techniques tested. Although two different sizes of training datasets were used, the accuracies of both models based on the normalization methods were similar. One of the future works will be to test the performance of LR algorithm on other datasets whose independent variables are vary in magnitude or those whose target variables are a 2-class problem. We will also try to see how Linear Regression algorithm performs under different normalization methods.

7. References

de Melo, V. V., & Banzhaf, W. (2016). *Improving Logistic Regression Classification of Credit Approval with Features Constructed by Kaizen Programming.* Paper presented at the Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion.

Evans, J. R. (2016). *Business Analytics, 2e*. Boston, MA: Pearson.

Jain, Y. K., & Bhandare, S. K. (2011). Min max normalization based data perturbation method for privacy protection. *International Journal of Computer & Communication Technology, 2*(8), 45-50.

Jayalakshmi, T., & Santhakumaran, A. (2011). Statistical Normalization and Back Propagationfor Classification. *International Journal of Computer Theory and Engineering, 3*(1), 89.

Kamijo, K.-i., & Tanigawa, T. (1990). *Stock price pattern recognition-a recurrent neural network approach.* Paper presented at the Neural Networks, 1990., 1990 IJCNN International Joint Conference.

Sola, J., & Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science, 44*(3), 1464-1468.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*(1), 3-14.

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Górczak, K., Klamecka, K., Figlerowicz, M., & Siatkowski, I. (2015). The impact of normalization methods on RNA-Seq data analysis. *BioMed research international, 2015*.