**Predictive Analysis – Case study**

**Customer classification**

Akila Selvaraj

Masters of Data Science,

Bellevue University

DSC 630 - T301

October 02,2022

**Predictive Analysis – Case study**

**Customer classification**

**Introduction:**

***Problem being solved***

In this case study, Sam’s club dataset was used to predict the customers who were more likely to pay for the purchase using cheque. This analysis was important for the company to predict the customers who uses cheque payments; so that they could target those customers for marketing campaign to apply for Walmart credit card.

***Importance of solving the problem***

Cheque payment is considerably risky as it is highly susceptible to fraud and theft. There is high probability that customers who pays using cheque do not use credit cards and they are the ones who will more likely to apply for company’s own card – Walmart credit card.From this analysis, the team was able to decide if customer will pay using cheque or not. The assumption in this analysis was that customers who use cheque payment as non-credit users. Customers who have credit cards sometimes use cash payment if the purchase price is low. So, for this analysis the target was only for the type of customers who use cheque payment but not the cash.

***Data acquisition***

Sam's club database contains retails sales information gathered from sales at Sam's club stores. When customer pays visit to the store and purchases the item, store associate scans the member's Sam's club card and then he proceeds with scanning each item in the purchase list. Customers information and sales information will be entered during the purchase. For this project, team used customers and purchase public dataset from Sam’s club.

# **Methods and Results:**

# ***Data Preparation***

# Original dataset had lot of anomalies variables with multi collinearity variables, skewness, kurtosis, missing values, etc. Also, the original dataset had lot of variables that are not related to target variables. So, to identify the significant input variables multiple methods like principal components, decision tree, and stepwise regression were used.

In Principal components method, the idea is that the smaller number of principal components represents most of the variability in the data and the relationship with the target variable. Therefore, instead of using all the original features in the Sam’s club dataset, they could utilize only a subset of the principal components.

Decision tree is a methodology which is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. A decision tree algorithm was used to split dataset features through a cost function. The decision tree was grown before being optimized to remove branches that may use irrelevant features by a process called pruning.

Stepwise regression was used which is the step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. It involves adding or removing potential explanatory variables in succession and testing for statistical significance after each iteration.

Also, log normal transformation was done for selective variables to eliminate outliers and achieve normal distribution.

## ***Data cleaning:***

Target variable in this dataset says if a customer used cheque for the payment or not. As it is a categorical variable, it was changed to binary variable with values 1 and 0 for easy model implementation. All the variables in the dataset were graphically analyzed and irrelevant data elements were eliminated from the input dataset. e.g., social security number, name, etc., that clearly have no effect on the target variable. As a part of this data cleaning, data was converted to an appropriate measurement scale, especially categorical (nominal scaled) data were converted to interval scaled when appropriate, eliminated variables with highly skewed distributions, eliminate inputs which were really target variables disguised as inputs, and imputed missing values. Similarly, income was represented as a character variable. The character may stand for $20K ($20,000), for $30K, etc., so the income variable was converted to an ordinal or interval scale, and a new version of the income variable was created in which all the values are numeric, and then eliminated the character version of income.

## **Problem solving**

***Modeling techniques used***

Neural networks is a classification model which tries to mimic the way the human brain develops classification rules and help us cluster and classify. Even with unbalanced nature of the sample distribution, neural networks makes it possible to generalize a successful model despite the irregularities in the data set. Neural networks model was applied on the dataset to perform the customer segmentations. Team also used Logistic Regression which performs binary classification by modeling a dependent variable in terms of one or more independent variables. This generalized model predicts the probability that an event will occur which in this case is, if a customer will use cheque for payment.

***Methods/Models used***

Team chose all these models to cross check which model provided lowest misclassification rate. Since the target variable was binary, misclassification rate was used as the primary method to calculate the best model.

PCA was used as it improves the performance of the model and it eliminates correlated variables that don’t contribute in any decision making. It also helps in overcoming data overfitting issues by decreasing the number of features.

Stepwise regression was primarily used to identify the significant variables necessary for the predictive analysis. Stepwise regression fits a logistic regression model in which the choice of predictive variables is carried out by an automatic forward stepwise procedure.  It involves adding or removing potential explanatory variables in succession and testing for statistical significance after each iteration. Stepwise regression circumvents the computational burden of trying all possible combinations of explanatory variables, by testing variables, one by one, in each step. The use of forward-selection stepwise regression for identifying the most statistically significant explanatory variables requires only a smaller number of regressions.

Logistic Regression was applied as it is easy to implement yet provides great training efficiency and this algorithm doesn't require high computation power.

***Metrics:***

Misclassification rate was the metrics used in this case study which tells the fraction of wrong predictions, without distinguishing between positive and negative predictions. In this case study, target variable is a binary variable and misclassification rate metric is easy to interpret and it is easy to calculate, so misclassification rate was used to identify which is the best model.

# **Conclusion:**

***Results and model implementation***

In the model, customers who would most likely to pay using cheque payment were predicted by using significant variables Member Type, total visit amount, sales tax amount, store number, and total unique item count. These variables were picked and found statistically significant with p-value < 5.

Baseline method is a method that uses heuristics, simple summary statistics, randomness, or machine learning to create predictions for a dataset. Based on target variable distribution, misclassification rate which is a performance metric that tells the fraction of the predictions that were wrong, was calculated. Then the misclassification rate calculated by the regression model was compared against other models and also against baseline model. Based on lowest misclassification rate, logistic regression model was selected as the best model.

***Actionable consequences of the case study***

Based on this case study, team were able to predict the customers who will most likely to pay the purchase using cheque. As a result of this case study, the company was able to identify and target those customers who will likely make payment by cheque and will send them marketing campaign for applying Walmart credit card.

***Learning from case study***

This segmentation case study allows the company to learn about their customers. They gain better understanding of customers and their needs and wants and therefore can tailor campaigns to customer segments most likely to purchase their credit cards.

***Approaching the problem differently in the future***

The input data set did not have much information about the customers. It had only the store and purchase information. If more variables related to customers like age, residential zip code, etc. were available, then the model could predict with even more accuracy i.e., with low misclassification rate.