Capstone Project – Humana Competition Case

Tanmoy Kanti Kumar

Supervisor: Dr. Murali Shanker

Department of Information Systems

Kent State University

Date 05-May-2021

Table of Contents

Introduction………………………………………………………………………………… 3

Project Goal…..……………………………………………………………………… 4

Data Exploration………………………………………………………….………… 5

Data Preparation…………………………………………………………………….... 6

Prediction modelling

Modeling Strategy ……………………………………………………………… 6

Brief Information About Feature Selection ………………………………………... 7

Evaluation of Model’s Performance……………………………………………... 9

Prediction on Test Dataset………………………………………………………….. 12

Introduction

Humana is a leading health care company that offers a wide range of insurance products and health and wellness services.

Social Determinants of Health (SDoH) are a key component of Humana's integrated value-based health ecosystem. 60% of what creates health has to do with the interplay between our socio-economic and community environments and lifestyle behaviors. Humana is seeking that "broader view" of its members to better understand the whole person and to assist them in new ways towards achieving their best health.

In the absence of regular, universal screening for SDoH, Humana needs to utilize robust data and advanced data science to understand which of our members are struggling with SDoH. This analysis will focus only on Transportation Challenges which is one of the major factors of SDoH.

Transportation Issue

Indicating the healthcare member who is having transportation challenges. A member having transportation challenges will not be able to get the proper healthcare needs. Predicting early of those members can help Humana to take necessary actions to mitigate their challenges and it will improve overall wellbeing of the member. Risks of having severe impacts can be predicted in advance.

About data

The data is provided by Humana. Data holds different parameter values of the healthcare members. Features are broadly classified into –

- Medical Claims Feature
- Pharmacy Claims Feature
- Lab Claims Feature
- Demographics/Consumer Data
- Credit Data

- Condition Related Features
- CMS Features
- Other Features

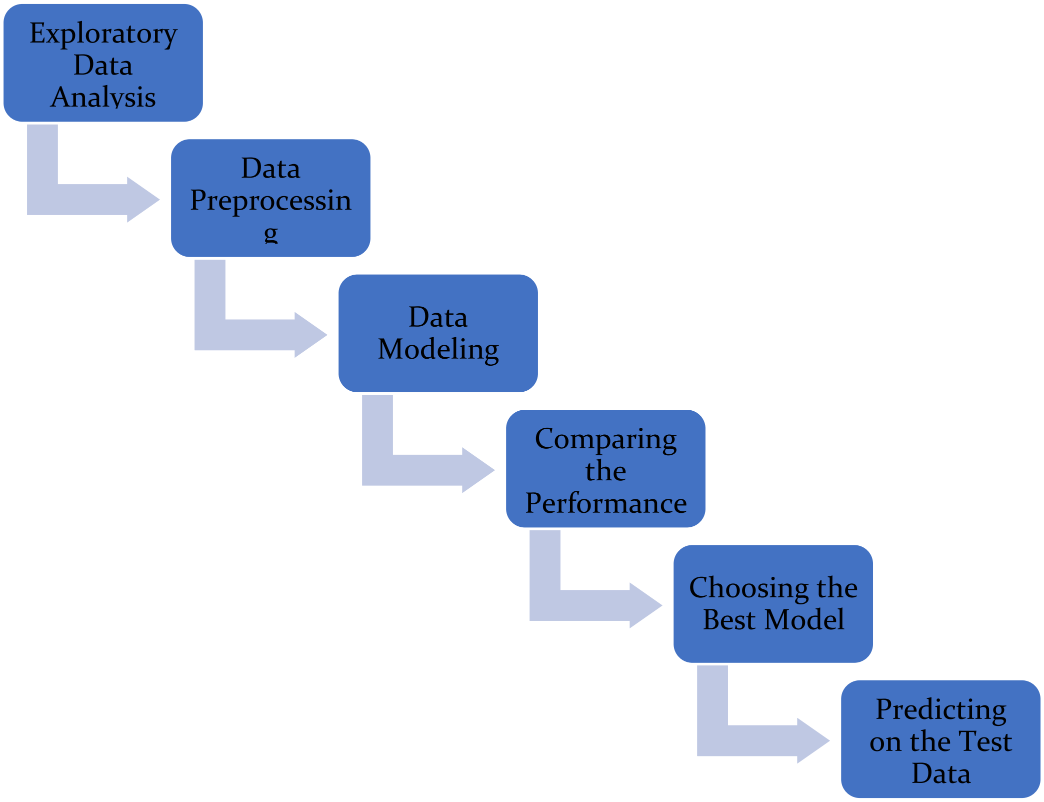
Project Goal

Predictive model -Since screening all Medicare members is challenging, having an effective predictive model to accurately identify members most likely struggling with Transportation Challenges is valuable. Data is provided and can be supplemented with publicly available data.

Proposed solutions–It is likely that members struggling with Transportation Challenges are not homogeneous and hence there are perhaps different solutions for different segments of members.

Process Flowchart

This High-level Process flowchart depicts sequential steps taken for creating the model.


Exploratory Data Analysis
Data Preprocessin g
Data Modeling
Comparing the Performance
Choosing the Best Model
Predicting on the Test Data

Data Exploration

Overview of Data

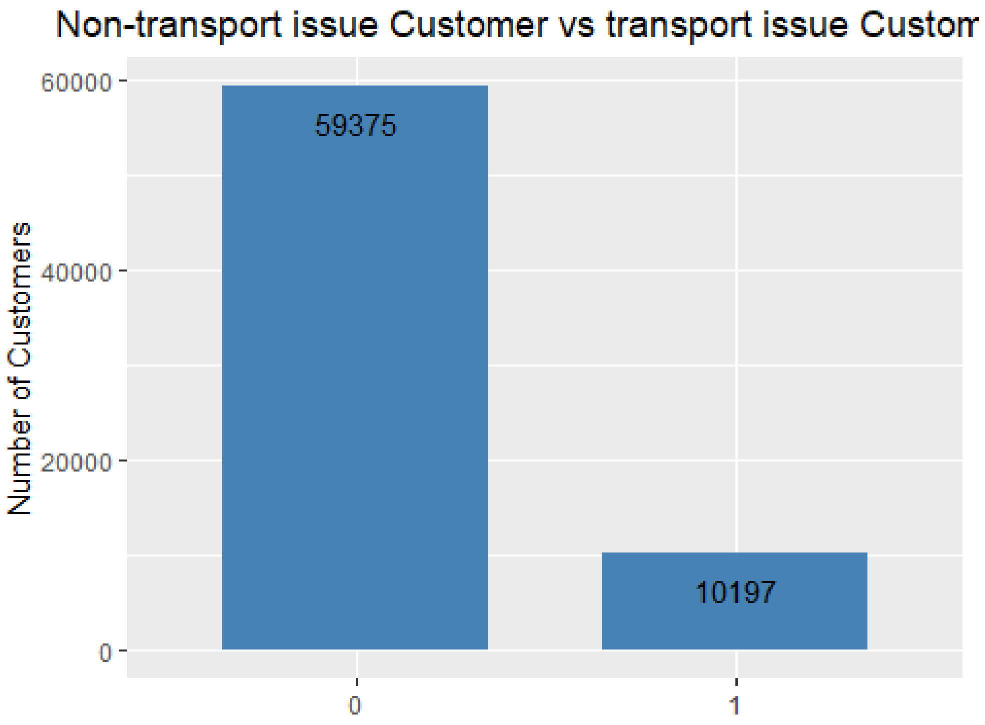
There are 69572 observations and 826 variables in the data set, character and numeric in nature. Out of the 826 variables first observation is holding unique customer/member id information and the last one holds the indicator whether the member has transportation issue or not.

Missing Values

The dataset did have some missing values (NAs). The amount of data missing from different observations ranged between 0% and 15.37% for each observation. We also found that the missing values for a specific variable can vary anywhere from 0% to 99.66% for each unknown variable.

Comparing Defaulters vs Non- Defaulters

In the training data, out of 69,572 customers, 10197 customers (17.17%) have transportation issue.



Data Preparation

A new binary variable 'transportation issue' has been made to factors for the classification model prediction.

- The first variable contained the unique identifier (Id) for every observation.
- The last variable covered if a customer had transportation issue.
- The remaining 760 variables were different attributes for the customers/members.

Part 1: Predicting transportation issue

Modeling Strategy

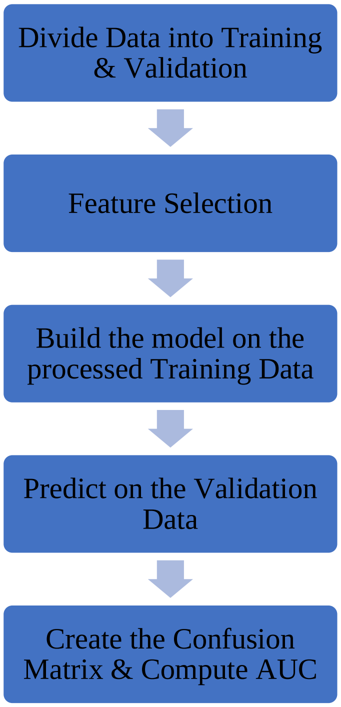
The priority was to manage the number of variables in the dataset. The dataset took quite a while to load in the R Studio. If I had to create a model on it, we had to reduce the number of variables so that it could run on limited resources in a fair amount of time. To achieve this, we used three different methods of feature selection:

- Removing highly correlated and near zero-variance variables
- Lasso (Least Absolute Shrinkage and Selection Operator)
- PCA (Principal Component Analysis)

Using these three methods, we ran below mentioned three classification models separately for each type of feature selection.

- Random Forest: This model builds multiple decision trees and merges them together to get a more accurate and stable prediction.
- Elastic Net: Elastic net is a penalized linear regression model that includes both the L1 and L2 penalties during training.
- Logistic Regression: This model in its basic form uses a logistic function to model a binary dependent variable.

The following process was followed to create and compare the models:


Divide Data into Training & Validation
Feature Selection
Build the model on the processed Training Data
Predict on the Validation Data
Create the Confusion Matrix & Compute AUC

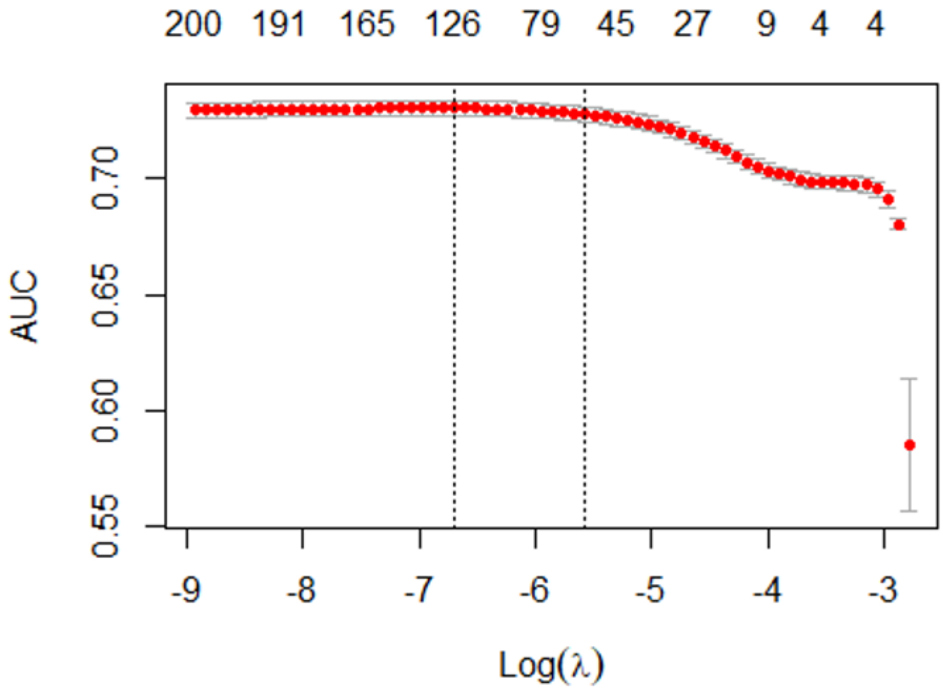
Brief Information About Feature Selection

Removing highly correlated and near zero-variance variables:

Near zero-variance variables were removed first. It reduced the number of variables from 826 to 404. Correlation among variables was computed next. The threshold was set to 50%, meaning that variables with correlation higher than 50% would be removed. It reduced the number of variables from 404 to 238. Missing values were imputed from the median of each variable.

Only LASSO:

The target variable was set as the binary variable that classified each observation on whether it had trs. issue or not. We used the lasso regression analysis method to perform both the variable selection and regularization, so that the bank could increase the prediction accuracy and interpretability of the people's default rates. The input for this model was 238 variables. The LASSO is a regression method that is used when people need to penalize the absolute size of the regression coefficients. In this project, the bank wants to know for each customer if they are approved or rejected by the bank based on the customer's default history. The below graph shows the Optimum Lamda value for the feature selection that was reduced from 238 variables to 126 variables. The first vertical dashed line indicates the lambda. Min value, while the second dash line shows the lambda value within one standard deviation to further reduce the variables. We decided to go with the most important 20 variables.



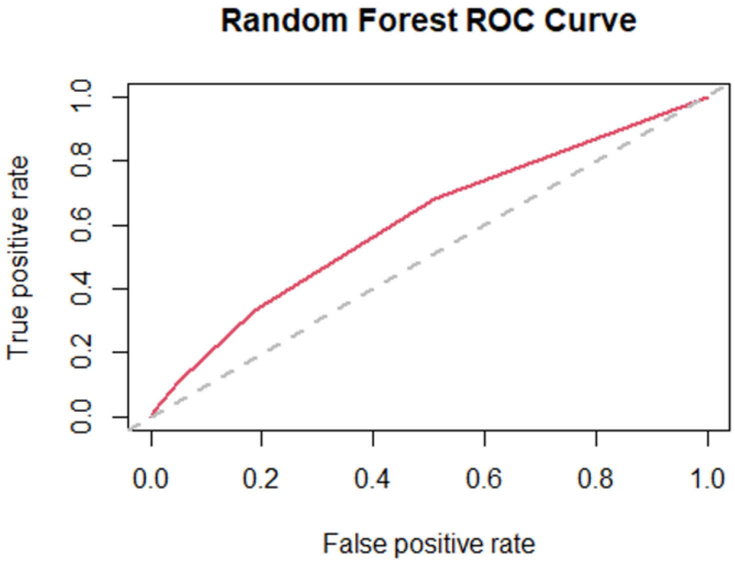
PCA

Selected 20 variables from the Lasso model and the Principal Component Analysis (PCA) was run on the remaining variables.

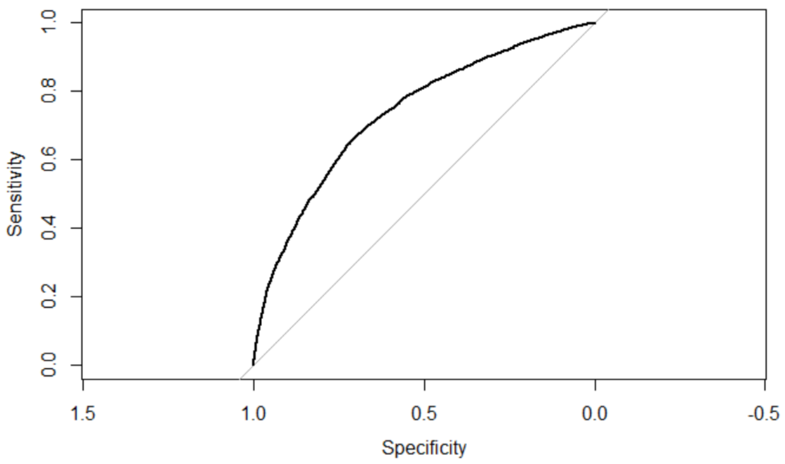
The PCA was used to reduce the remaining large dataset of variables into an even smaller amount to better handle the information but at the same time holds the most valuable piece of information in the process. PCA reduced the number of variables to 69. The 69 components captured 75% of the variance.

Evaluation of Model's Performance:

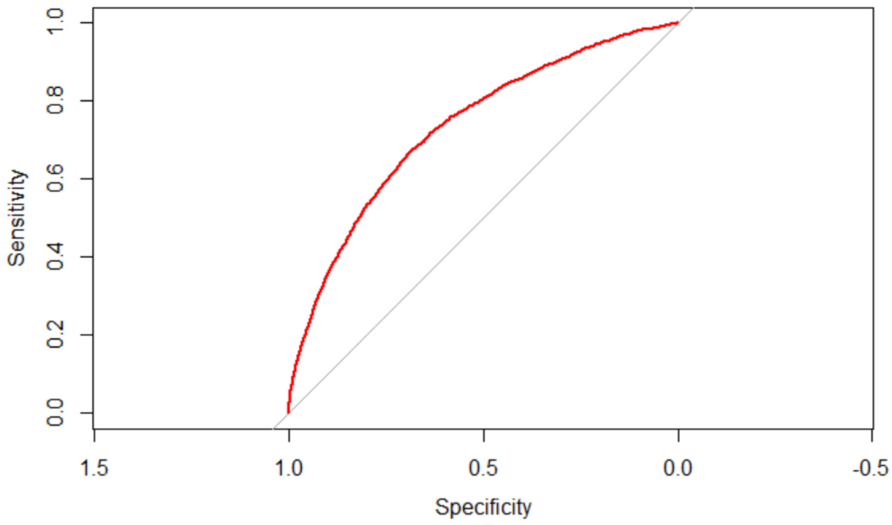
Used AUC as a metric to measure model performance. This process takes into account the True Positive and False Positive probabilities when samples are chosen from the prediction model. We compared the AUC and the accuracy of all the 9 models which we built using different feature selection methods and we chose the model with highest AUC value to predict the probability of default on the test data set.



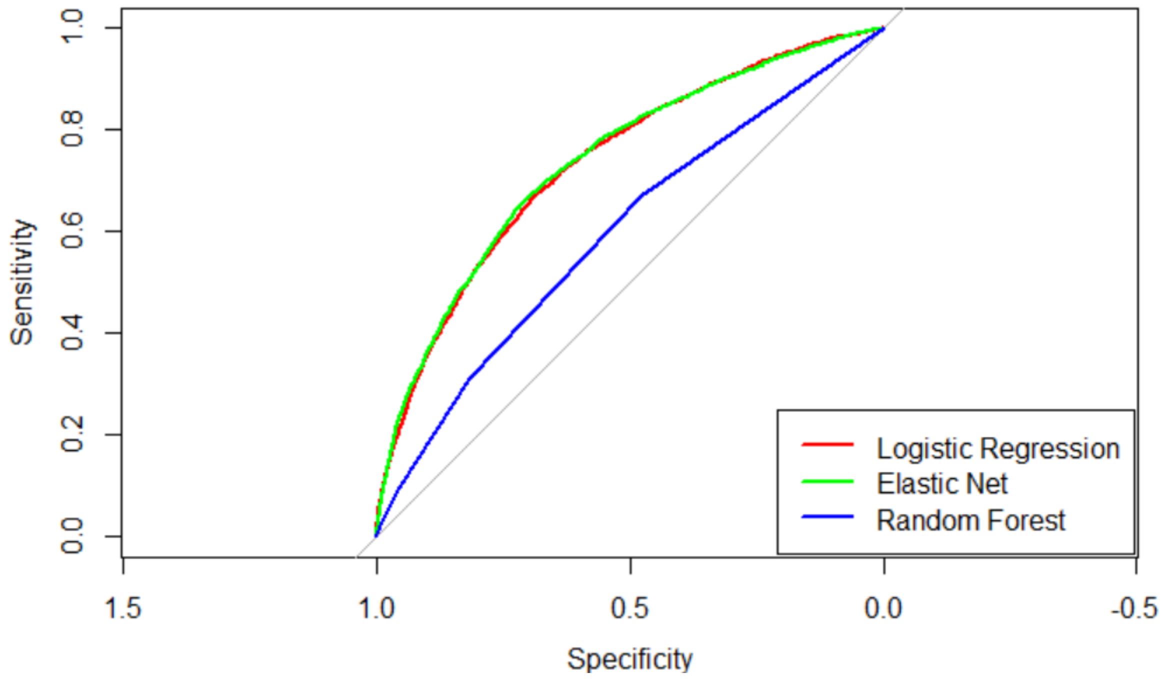
Elastic Net:



Logistic Regression:



As we can see from the above graphs, the Logistic Regression models works best in all the three scenarios. I have choosen the best model based on the highest AUC value.

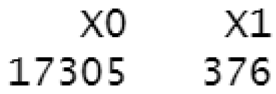


Models	**AUC**
Random Forest	0.6087
ElasticNet	0.7359
Logistic Regression	**0.7354**

The Logistic Regression model with PCA & LASSO was chosen as it had the highest AUC.

Prediction on Test Dataset

The test dataset also had missing values which were imputed from the median of each variable. Also, it was preprocessed with PCA and LASSO in order to have the same number of variables as that of the train data set. Then I used the Logistic regression model to classify the test dataset and below were the results.



```
   X0      X1
17305     376
```

The number of customers predicted to have transportation issue 376.