Amex Credit Risk Analysis Model

Himasree Viswanadhapalli

December 2021

1 Introduction

A credit risk model that predicts the likelihood of a customer to default a payment is developed. The data present for a customer is at any given point of time and a model is developed to predict the likelihood of the customer to default after 12 months. The training data set contains 83 thousand observations and test data set has 47 thousand observation that needs to be scored by the developed model. The data set has the customer application and bureau data with the default tagging i.e., if a customer has missed a cumulative of 3 payments across all open trades, his default indicator is 1 else 0. Data consists of independent variables at the time T0 and the actual performance of the individual (Default/ Non Default) after 12 months i.e., at time T12. 49 features are given in the data set from 'mvar1' to 'mvar47' including application key and default ind (indicator) features.
The aim of the project is to develop a machine learning model for customer default prediction. CatBoost model and Random Forest model probabilistic predictions are stacked with corresponding weights to obtain a best performing model after various feature engineering techniques.

2 Data Visualization

After looking at the data, few observations are made regarding the missing values. Features having missing values are represented by NaN, 'na', and 'missing' in the data set. However, 'missing' and 'na' are not considered as missing values but they are considered as a separate category. The features with 'missing' and 'na' values are considered to be object which needs to be converted to int or float as all the other entries are numerical. To handle this, the 'missing' and 'na' terms needs to replaced by NaN followed by handling missing values. Figure 1. represents the percentage of missing values in the data corresponding to each feature. New features are not created as the given features are not explained.

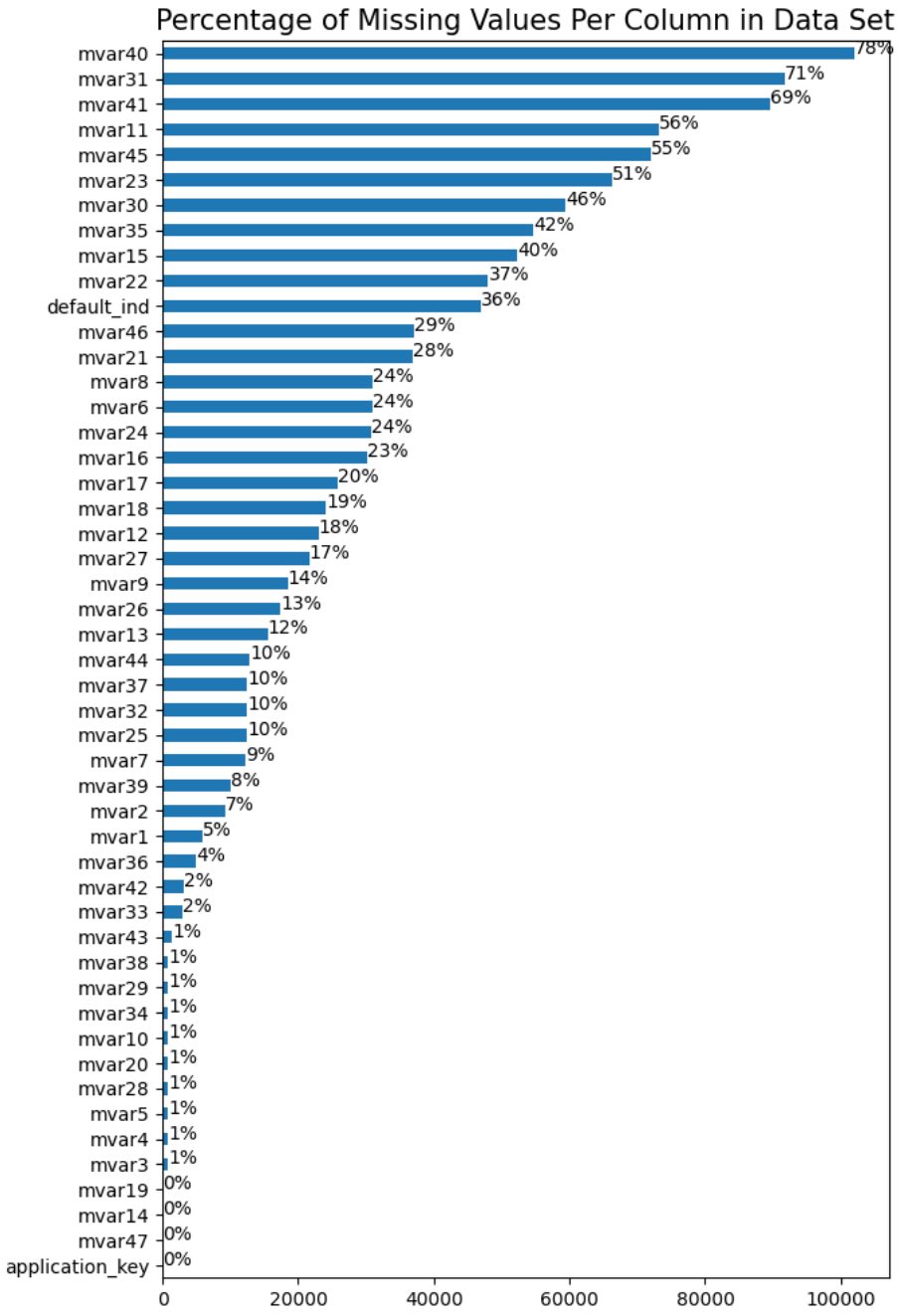

Figure 1: Visualization of Missing values

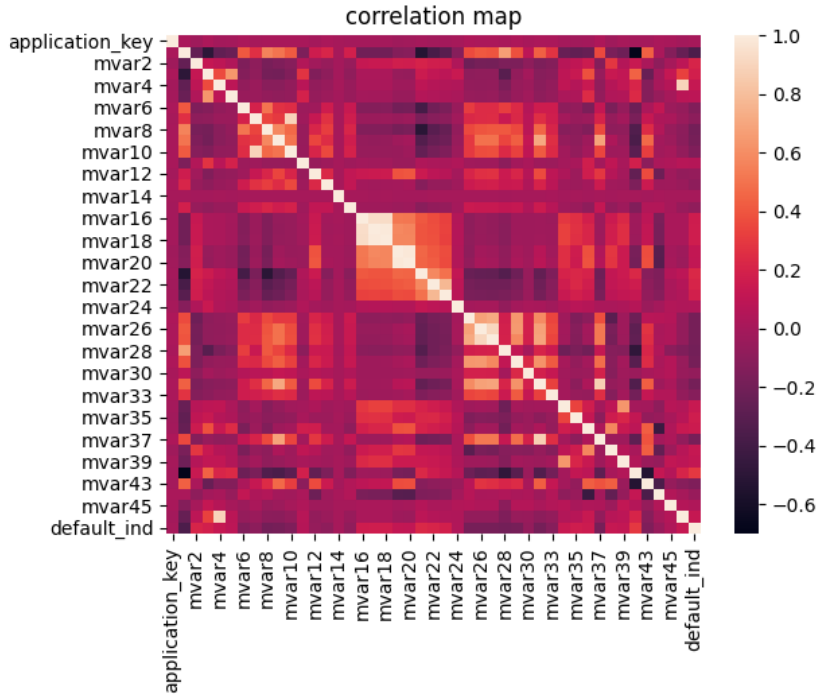

Figure 2: Correlation Map

3 Feature Engineering

3.1 Handling Missing Values

Features with high missing values, i.e., greater than 60% missing values are dropped. The features are divided into numerical and categorical features to handle the missing values separately. The numerical features are further divided into continuous numerical features and discrete numerical features. The distributions of numerical features with missing values is visualized. Only one feature 'mvar1' has a almost normal distribution and all other features follows skewed distribution. As the distributions are skewed, the mean can not be used a central tendency to fill the missing values. Therefore median has been used to fill the missing values of continuous numerical features. However for 'mvar1' either mean or median can be used as both the values are observed to be close to each other. In case of categorical features, their are no categorical features with missing values.

3.2 Feature Encoding

3.2.1 Encoding of Categorical Features

The categorical features are converted to numerical input before the model development using different encoding techniques. There are different encoding techniques like one hot encoding, count encoding, mean encoding, target guided count encoding, target guided mean encoding, KDD orange technique, etc. The categorical features are encoded. The only categorical feature in the data set is 'mvar47'. Count based encoding is used to encode the feature. The feature has two categories 'C' and 'L' where the frequency of 'C' category is high. Therefore the 'C' category is encoded into 1 and 'L' category is encoded into 0.

3.2.2 Encoding of Discrete Numerical Features

The discrete numerical features are also encoded after visualizing the features. The numerical features with number of unique values less than 100 are considered as discrete features. None of the discrete numerical features follow the normal distribution. Therefore encoding the discrete features by converting them into categorical features improve the model performance.
The categories with frequency less than 2.5% in the discrete numerical features are converted to single category 'Rare Var' which indicates that the rare categories in the feature. In the next step after handling the rare categories, the discrete numerical features are divided into low count features and high count features. low count features represents the features with number of unique values less than 10 and high count features represent the features with number of unique values greater than 10. Count based encoding is used to encode the low count features where the categories are ranked based on frequency of occurrence. The high count features are encoded using target guided mean encoding technique. In target guided mean encoding, the categories are ranked based on mean of target variable corresponding to each category on the feature.

3.3 Standardization

The data is standardized using standard scalar from sklearn to arrange the data in standard normal distribution. In standard scalar the data is arranged with mean 0 and standard deviation 1. The mean is subtracted from the data points and scaled to unit variance by dividing with the standard deviation.

4 Model Development

The first step in training the model is to split the data set into train set and test set. The data is split with 10% test size. To develop the best model, different machine learning models like XGBoost, Random Forest, CatBoost, logistic regression has been trained and hyper parameter tuning is performed for each of the models for better performance. Randomized grid search technique

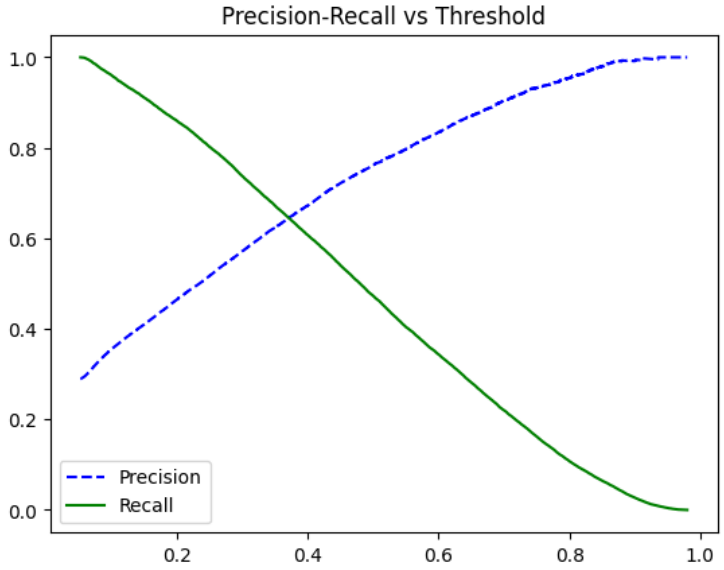

Figure 3: Precision Recall curve at different Thresholds

has been used for the hyper parameter tuning. CatBoost over performed the other models in terms of F1 score. The CatBoost model is trained ROC AUC score as the evaluation metric which improved the performance.

4.1 Threshold Calculation

Two types of threshold values are used to classify the prediction probabilities into defaulters and non defaulters. The first type of threshold is Youdens cut-offs where the threshold maximizes the ROC AUC score, the second type of threshold maximises the F1 score of the model. For calculating the Youdens threshold value the true positive rate and false positive rate are computed at different thresholds using roc_curve function from sklearn.metrics.For maximising the f1score, recall and precision values are computed are different threshold values from precision_recall_curve function.Figure 3 represents Precision Recall curves vs Threshold. To get the threshold values the ROC AUC values and F1 score values are computed at different threshold and the threshold values are sorted based on the scores. The best threshold value is the value with best ROC AUC score or F1 score. For the purpose of this model, The threshold which maximises the F1 score is considered and it is also observed to improve the performance of the model.

4.2 Stacking

The models are stacked with weights which sum up to 1. The CatBoost model again out performed all the stacked models. Hence, the CatBoost model is finalized as the best performing model. A Catboost model and random forest model with weights 0.9 and 0.1 respectively are stacked to produce the result of 60.67% on the American Express portal. While stacking the predicted probabilities are multiplied with weights and added to get the weighted probabilities. Weighted probabilities are further used to calculate the threshold value and for final classification.

5 Other Techniques Used

Number of other techniques are used in model development which did not improve the model performance significantly.Firstly, there are several uni-variate outliers in the data set which are either removed or replaced with central tendency. Multivariate outliers are detected with help of Mahalanobis distance. The Mahalanobis distance is calculated for both the test data set and train data set. The data points with Mahalanobis distance out of range of Mahalanobis distance of test dataset are dropped. Secondly, the features with high multi collinearity are dropped as they induce variance in the coefficients of the model, however this technique dropped the performance of the model. The collinearity is calculated with the help of corr function from seaborn. Figure 2 represents the correlation heat map of the data. Thirdly, neural networks model is developed, due to the presence of high missing values which are replaced in the data set as well as due to the presence of high number of outliers and imbalanced data set, neural networks did not out perform the machine learning models. SMOTE technique is also used where defaulters class is over sampled and a given ratio (r=0.4,0.5,1) of number of defaulters and non defaulters are generated.
Finaly, the logit regression is used to get the predictive power of features and Akaike information criterion (AIC) value. Lower AIC indicates that the model is better. The features with high p values are selected and different encoding techniques has been used to reduce the p value to 0 and to decrease the AIC value. However with all the feature p values approaching zero and lowest AIC value of all the models, the model performance did not increase significantly. The encoding techniques used to reduce the AIC value are discussed in the next section. In one of the approach, the density plots of the features are plotted for both default and non default categories. The features with same distributions for both defaulter's and non defaulters are dropped as the model gets confused with these features. This technique worked for other machine learning hackathons, however the model under performed for the given data set.

5.1 Techniques to Reduce the AIC Value

Different encoding techniques are used for categorical features and numerical features. Considering the numerical features, if the corresponding p value is high, the feature is converted to categorical features. If the number of unique values are high in the feature, the rare categories are handled which is discussed in previous sections and the encoding process is repeated. if the number of unique values are low or if the feature is categorical feature, either label encoding or one hot encoding technique is used. After following the mentioned encoding process, the p values co responding to the selected features approached zero and the AIC value reduced.