Data Wizards (Group 4) Project 2

Di Chen Mai Castellano Tyler Kussee Spencer (Hutchison) Yang

Introduction

In our pursuit of statistical inquiry, we have chosen to explore the Vehicle Loan Default dataset, comprising approximately 41 columns, with one designated as the response variable. Encompassing diverse information, the dataset delves into loan details, including date of birth, employment type, and credit score, alongside loan-related specifics such as disbursal details and loan-to-value ratios. The dataset presents challenges, notably in the form of odd date and time length columns, requiring standardization and transformation into comprehensible formats conducive to model development.

We want to discover the most influential explanatory variables driving loan default, and their impact within the dataset. We also want to find the optimal modeling approach for harnessing the training data, evaluating various methodologies to identify the most effective. Ultimately, our investigation extends to which among them best identifies the underlying dynamics of vehicle loan default prediction.

Obtain/Scrub the data

The data was pulled from the Vehicle Loan Default Prediction datasets available on Kaggle. As mentioned earlier, we have approximately 41 columns, with one of the columns designated as the response variable. First, we'll import the dataset from the training CSV:

We then scrub the unknown values in our length and age fields:

Then we do our calculations and create various other fields for our modeling usage:

Now that we've combed through our dataset, we can now going through our dataset to extract our features and learn the optimal model for classification.

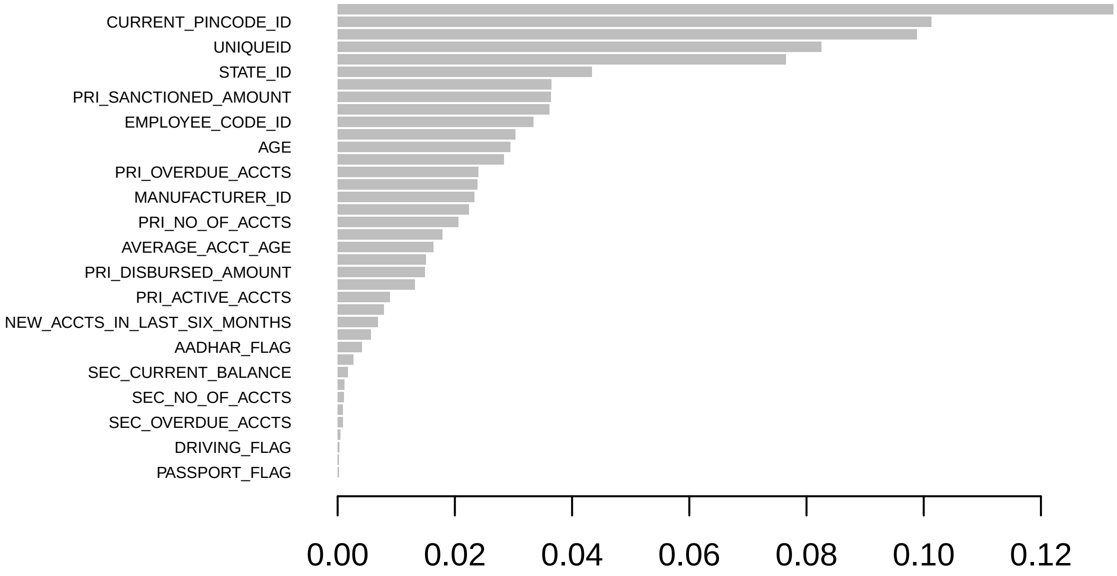
Explore the data

So in the case of this dataset, we need to find the features that have the biggest impact on our "LOAN_DEFAULT" variable. In order to find this, we will be using the XGBoost machine learning algorithm package to train our matrix and label vector from the training data, setting the parameters, and performing cross-validation to find the optimal number of rounds for training the model. Once the parameters and optimal number of rounds were found, we trained the model and then proceeded to calculate the feature importance scores. From there, we select the top 10 features based on their importance to the model, along with plotting the scores to visualize the relative importance of each feature. With these scores and our graph, we found the following features to be the most important:

- LTV
- CURRENT_PINCODE_ID
- PERFORM_CNS_SCORE
- UNIQUEID
- DISBURSED_AMOUNT

- STATE_ID
- SUPPLIER_ID
- PRI_SANCTIONED_AMOUNT
- AGE
- EMPLOYEE_CODE_ID

With all of this in mind, we're able to now subset the training dataset with these important features, and split it into new training and testing datasets. we can start building our classifiers and getting results.



```r
# Subset the dataset with the selected features
filteredData <- train[, c(topFeatures, "LOAN_DEFAULT")]

# Training/Testing split
trainIndex <- createDataPartition(filteredData$LOAN_DEFAULT, p = 0.8, list = FALSE)
trainData <- filteredData[trainIndex, ]
testData <- filteredData[-trainIndex, ]
```

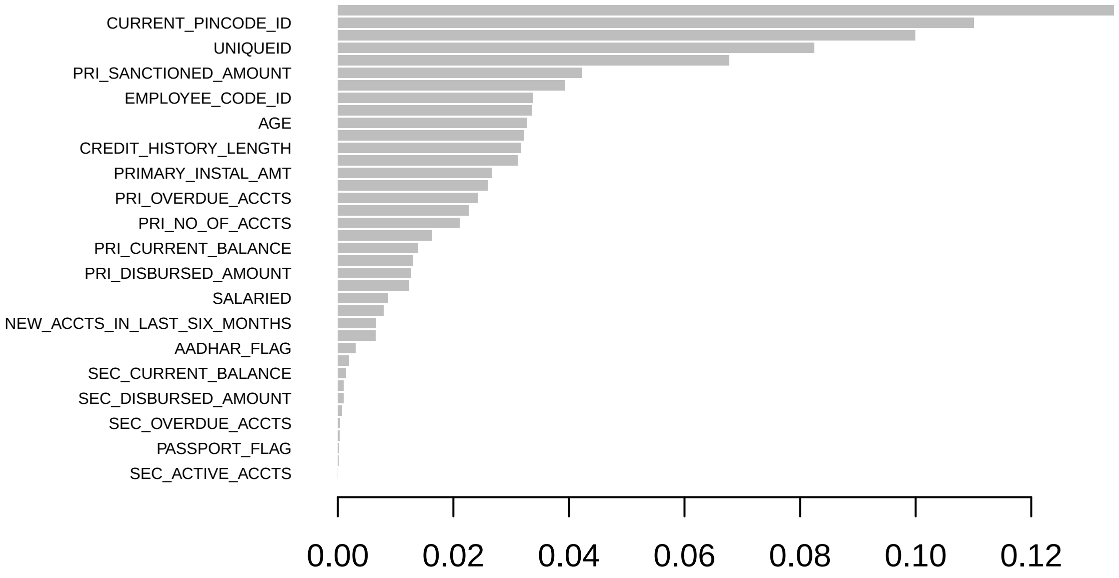
Model the data

Interpret the data

Obstacles

Conclusion

Appendix



```r
# Subset the dataset with the selected features
filteredData <- train[, c(topFeatures, "LOAN_DEFAULT")]

# Split the selected data into training and testing sets
trainIndex <- createDataPartition(filteredData$LOAN_DEFAULT, p = 0.8, list = FALSE)
trainData <- filteredData[trainIndex, ]
testData <- filteredData[-trainIndex, ]
```