Classification Project Writeup

By: Divya Uppal

April 2022

Abstract:

The goal of this project was to identify customers who will default on credit card payment by developing a predictive classification model. Data was gathered from UCI Machine Learning Repository and had a combination of numerical and categorical features. Random Forest was used to achieve promising results to identify high risk debtors.

Impact: Non payment of credit by customers diminishes asset quality, profitability of credit card companies. It also decreases future capacity of credit sanctioning. Measures such as contacting the borrower with the aim of re-establishing some form of regular payment schedule; using an outside collection agency etc are performed to collect outstanding funds.

Design:

Data:

Project data was gathered from UCI Machine Learning Repository and had 30K Data points and 23 features. Features were a combination of categorical and numerical variables. Few feature highlights include payments made during the last 6 months, amount of credit available, highest education, gender, age etc. Interaction variables were also created to identify relationships between features. 9 most important features were identified which determined the default probability of a customer.

Algorithm:

Feature Engineering:

1. Converting categorical variables to dummy variables
2. Creating interaction variables using dummy variables and numerical features
3. EDA to manage missing or un-identified categories

Models:

Logistic Regression, k nearest neighbor and Random Forest were evaluated before settling for Random Forest with highest cross validation performance. Random forest was also used to perform feature selection and identify important features which predict probability of default in customers.

Model Evaluation & Selection:

- Data with 30k data points were divided into training/test dataset (80%/20%)

- Hyperparameters were identified using 10 fold cross validations on training data
- Threshold while making hard classification was also identified by evaluating performance on Validation dataset
- Class Weights were used to deal with imbalanced data set
- Threshold while making hard classification was used to manage imbalance in the data set
- F2 Score was used as a metric for evaluation. Recall was given twice as high as importance in comparison to precision

Final Model:

Random forest with 10 fold cross validation: 37 features, Class Weight: Balanced, No of estimators = 300, Max Depth = 20

F2 Score: .6028

Accuracy: .6555

Confusion Matrix:

2972 (True Negative)	1701 (False Positive)
366 (False Negative)	961 (True Positive)

Precision = .36
Recall = .72
F2 Score = .60

Tools:

- Scikit-learn for modeling
- Numpy and Pandas for data manipulation
- Seaborn and Matplotlib for plotting