Case Study

**Chase’s Prediction of Mortgage Risks**

DSC 630 Predictive Analytics

Ganesh Kale

**Overview:**

This case study describes an analysis performed on Chase Banks Mortgage Risks Project. The Chase Bank is an American national bank and offers different services such as Personal banking, Home Loans, Credit cards etc. One of the services they offer is Home Loans that is Providing mortgage to customers with competitive APR. Chase bank earns around 43% of revenue from Home Loans so one of the goals of bank is to increase the revenue year over year and reduce the risk in Mortgage portfolio.

Chase’s mortgage portfolio faced risk factors amounting to hundreds of millions of dollars. There are couple of key Mortgage risks such as - 1. Customer defaults on Mortgage Payment and 2. Mortgage early prepayment. Both the risks cause reduction in home loans profit.

Chase bank holds all the information about customers when they offer mortgage loans and with use of these data, they wanted to predict the customers who get defaults in the cycle of loan repayment. With help of the prediction Model bank would flag the customers who might get defaults and offer appropriate loan repayment plans or any other promotion to avoid delinquency. This prediction model could save bank some millions of dollars annually.

**Business Understanding:**

Loan Origination is a platform in Chase bank where they offer home loans to customers based on their credit score, income, employment etc. When customer apply for the home loans, they need fill in all the personal information, educational information, financial information and based on this chase approve loan with certain APR. Also chase possesses the loan servicing information when customer start repaying the loan such as number of payments made, new credit inquiries, credit card balance etc.

For this project chase had all the required information they needed to build the predictive model, additional data such as credit card payments, balance, checking or saving account balance etc. retrieved for the customers who were having account with chase bank.

Problem Statement – Identify the potential customers who will default on loan payments.

Such customers are risk to mortgage portfolio so identify them early in loan servicing will be helpful so such loans can be sold out or customers will be offered with appropriate offers.

To build the predictive model based on historical data wherein customers with delinquent and customers who closed loans need to be collected and the feature that states loan status either delinquent or closed will be target variable.

Data collected for this PA model will be having balanced data on target variables to avoid bias in decision making.

**Data Understanding:**

For this project 50K loans are identified with equal combination of loans with closed and delinquent statuses with different combinations of impacting features such as credit score, income, employment, education, number of payments and late payments etc.

Below table shows the key features with description that are used for building predictive model.

**Data Description:**

|  |  |
| --- | --- |
| Features | Description |
| customer Id | Customer Unique Identifier |
| credit score | Credit score of the customer |
| income | Customers net income |
| employment Status | Current employment status |
| highest education | Customers highest education level |
| Assets | Customers new assets other than income, savings, CDs etc. |
| Credit Card Balance | Credit Card unpaid balance |
| CC late Payments | Credit card late payments |
| Payments made | Number of on time mortgage payments |
| Missed Payments | Number of missed mortgage payments |
| late payments | Number of late payments |
| Loan amount | loan amount |
| property value | Home value |
| property type | Home type single or multifamily etc. |
| age | Age of customer |
| marital status | Marital status |
| Other Loans | Customer having other loans Y/N |
| loan status | delinquent or closed |

Since the data collected in house (from chase) so there are very minimal chances of having data discrepancies.

**Data Preparation:**

The data collected for this project having numerical and categorical data types. The target variable here ‘loan status’ is also categorical data. To prepare data for modeling stage, we need to transform all the features into numerical.

Missing values – Since data collected from Chase’s historical loan data so most of the features were having values. Only the concern with features of which data was pulled from chase banking such as credit card balance, assets etc. will be null for customers who do not have accounts with chase. To fix this issue, the data is filled with values from the application when customer applied for the loan.

Data Transformation – Most of the data is categorical or Boolean, we need to transform them in to numerical for modeling purpose. Here, all the features with Boolean values are transformed to 0’s for False/No and 1’s for True/Yes. The categorical variables are transformed to numerical using pandas ‘get\_dummies’ method.

Feature Correlations: Since the problem statement is to identify delinquent customers, we need to check the multi-collinearity among all the predictors and correlation with target variable. This exercise helped to identify features which are strongly correlated with target variable and potentially affect the loan status. This is performed using visualization and pandas corr method.

**Modeling:**

After performing the EDA and understanding the features that are strongly correlated with target variable, we need to see how to use these key features to predict customers loan status by training data using different machine learning algorithms.

First step followed in modeling is to split the data into training, and testing. With help of scikit-learn package we divided data into 70% and 30% for training and testing respectively.

The first model used for this project is logistic regression, since the target variable is categorical data type and is binary variable so logistic regression is best and simple algorithm to train model to predict customer loan status (default or closed). The logistic regression is supervised ML technique used to predict the probability of certain class or event, since our data has the binary target variable, so it is the best algorithm to start with. The logistic regression gives the output after training on train data, the value of r-squared and adjusted r-squared are the key values, the value close to 1 or 100% signifies what percentage of the factors accounts for predicting the target variable. Also, the coefficient of each feature denotes how much it affects or impacts target variable when a unit change in that variable assuming all other factors constant.

The second algorithm used to train model is XGBoost, the XGBoost algorithm is effective for a wide range of regression and classification predictive modeling problems. It is an efficient implementation of the stochastic gradient boosting algorithm and offers a range of hyperparameters that give fine-grained control over the model training procedure. Since predicting customer will be default or close the loan is a type of classification problem so this algorithm should be best fit for this model.

After training data on both algorithms, the models are tested using test data and predictions are made, then accuracy score is calculated to compare the models. The accuracy score tells us which model can predict a most number of target variable correctly. Based on the comparison XGBoost model accuracy score is better than logistic regression.

Model trained using XGBoost algorithm is the baseline model and will be used in prediction model.

**Deployment:**

The baseline prediction model that would predict the customer would default on payment or close the loan is ready for deployment and test on real time data. To make sure model prediction is accurate, all the loans that flagged by model will be logged and assessed by SMEs to make sure model predicting accurate loans. This verification will be done for initial period and them model will be full-fledged used on real time data.

The model performance and logs will be monitored on timely manner to made sure model is working and producing the result what it supposed to produce that is flagging customer accurately whether they will default or close loan.

Also, model will be trained on latest data to make sure all the new scenarios and new features considered when predicting the customer loan status.

**Summary and Conclusion:**

To resolve chase banks mortgage risk problem, we have built the predictive model to predict the customer who would default, or close loan based on certain factors. Here two key algorithms are used to train and test and the model with highest accuracy score is baselined. This kind of approach is helpful for bank since bank does not need to invest on getting data to build model and model built using in house data will be more reliable on in house data.

The final model will help chase bank to identify potential customer who might get default on their loan payments so that chase can invest some to offer them loan repayment plans or any kind of promotion to keep customer for long term and close the loan. Though this model would not be 100% accurate but would predict 80% + accurate result and help automate the decisions and save some manual efforts of analyzing the loan for delinquency.

This model would definitely improve chase banks overall mortgage portfolio revenue and enhance customer experience.