CASE STUDY

BANKING DATA ANALYSIS

SYNOPSIS/PROPOSAL

Submitted By:

Anukriti (2K17/IT/027)

Anurag Mudgil (2K17/IT/029)

PROBLEM STATEMENT

A lender commonly makes two types of decisions: first, whether to grant credit to a new applicant, and second, how to deal with existing applicants, including whether to increase their credit limits. Credit scoring is the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit and taking these two decisions effectively. These techniques determine who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of the borrowers to the lenders. Further, they help to assess the risk in lending. Credit scoring is a dependable assessment of a person's credit worthiness since it is based on actual data. Our project has various application areas in the consumer market including credit cards, auto loans, home mortgages, home equity loans, mail catalog orders, and a wide variety of personal loan products.

PLAN OF EXECUTION (PROPOSAL):

In this project, we intend to address the decision making problem of the lenders by effectively analyzing a large sample of previous customers with their application details, behavioral patterns, and subsequent credit history available through which we will be able to identify the connection between the characteristics of the consumers (annual income, age, number of years in employment with their current employer, etc.) and their subsequent history. With the help of this, lenders will be able to predict the defaulters and decide whether to grant credit to a new applicant.

❖ COLLECTION OF DATA (Dataset to be used):

Data will be aggregated from 2 different sources as mentioned below:

- Give Me Some Credit (Kaggle dataset)
- Credit Risk Modeling Case Study (Kaggle dataset)

Description of the final dataset obtained:

The dataset consists of about 150k observations and has 13 variables. These variables can be broadly divided into these categories:

1. Demographic details: Age, number of dependents.
2. Financial status: Monthly income, debt ratio, utilization of credit lines etc.
3. Delinquency history: How many times the user has default in the loan payment in the 30-60 days, 60-90 days and more 90 days past the due date, Serious delinquent (loan defaulter)

Some of the variables are explained as follows:

- Revolving Utilization of unsecured loan: Ratio of the credit balance (credit used) and the total credit limit
- NumberOfTime30_59DaysPastdueNotW: Indicates how many times a customer failed to pay in 30 to 59 days past the due date of payment in the last two years. Similar logic is used for the computation of NumberOfTimes90DaysLate, NumberOfTime 60_89DaysPastDueNotW.
- Age
- Debt Ratio: Ratio of the all the loan payments to gross income
- Income
- No.of dependents
- Categorical: Gender, Occupation, Region, House Status and Education

❖ FEATURE ENGINEERING:

Feature/Data Transformation:

We will be using **Weight of Evidence binning method** where we will be first transforming all the independent variables (like age, income etc.) using the weight of evidence (WoE) method and based on the proportion of good applicants to bad applicants at each group level, this method attempts to find a monotonic relationship between the independent variables and the target variable.

For a continuous variable, we will be first splitting the bins around 10 (or 20) after which will calculate the % of Good events and % Bad events and finally, will replace the raw data with the calculated WOE values.

Feature Selection:

Some features might require more processing due to the presence of outliers or missing values. So, observing our dataset, a cross table will be created based on age group and income group where the median value of no. of dependents is calculated for each combination. The missing values in this variable are thus imputed based on the age and income of the customer. We will also be creating dummy variables for gender, occupation, region, house status and education.

T-test and VIF:

We will also get features importances by performing independent t-test and VIF on each variable and hence, checking whether all the variables are within the significance level and if not, we will be eliminating the unnecessary features before applying our classification algorithm. check to obtain best parameters and fitting them best parameters to obtain a better accuracy, if needed. VIF test will be applied to check multicollinearity between our variables based on which can further select our features.

❖ CHOICE OF MODEL/CLASSIFICATION ALGORITHM:

On the basis of study from (PDF) Credit scoring methods we arrived at these approaches :

1. Linear Discriminant Analysis : It is simple and easy to estimate but it is optimal only for small distributions and gives accurate results only for normally distributed data.
2. K nearest Classifier : It is a non parametric approach but there is a lack of a formal frame-work for choosing the k and accuracy decreases for high values of K.
3. Classification and Regression Trees : It is a flexible and potent technique but there is computational burden in case of large datasets. CART optimizes only locally on a single variable at a time and thus it may not minimize the overall costs of misclassification .
4. Neural Networks : Genetic algorithm (GA) and MultiLayer Perceptron are majorly used. They have high accuracy but the major drawback of these is their lack of explanation capability.
5. Logit Regression : It is an extension of the LDA model that allows for some parametric distribution. It has the advantage that one can provide the probabilities of being in some concrete class. The logit method can also deal with categorized data .
6. Random forest classifier: Since it predicts using the outcomes from various decision trees, it offers a lot of advantages. Firstly, allows us to quickly and easily change the output to a simple binary classification problem. Secondly, it allows us to output a probability score.

Both Neural Networks and Logit Regression are widely used approaches but on studying (PDF) Credit Scoring Models: Techniques and Issues we get Logit Regression has higher accuracy so it would be a better classifier model . Based on the above analysis, we came up with an Ensemble technique using Random forest classifier along with Logit Regression algorithm.

❖ PERFORMANCE METRICS:

Combination of below mentioned metrics:

- ❖ Accuracy
- ❖ Precision
- ❖ F1-Score
- ❖ Recall
- ❖ Area under ROC curve
- ❖ Gain Chart

TECHNOLOGY STACK:

- ❖ **MACHINE LEARNING CLASSIFIER:**
  - **Libraries:** pandas, itertools, bs4 (BeautifulSoup), sklearn (classifiers, metrics, model_selection), matplotlib (visualization) etc.
  - **Tools:** Anaconda, WEKA
  - **Language:** Python

- ❖ **GUI/DEMO:**
  - **Frontend:** React
  - **Backend:** Flask
  - **Languages:** Javascript, Python

GUI/DEMO:

We will be creating a simple web application where users can enter their demographic details like age, number of dependents etc, financial status (Monthly income, debt ratio, utilization of credit lines etc.) as well as their delinquency history, for instance, how many times the user has default in the loan payment in the 30-60 days, 60-90 days and more 90 days past the due date. This data will then pass through our model with the help of which we will be calculating the customer credit score. This score will finally help the lenders decide whether to grant credit to that particular customer. This way they will be able to take decisions more quickly as well as effectively.