Loan Decision for Borrower

1 Problem Statement

Build a predictive model that borrowers (credit/loan seekers) can use to help make the best financial decisions. This is predicting the probability that somebody (i.e. borrower) will experience financial distress in the next two years.

2 Explanation

This problem is mainly creating a system from load data to predict possible defaulter given any new customer. Here we have credit defaulter data where our label is customers defaulted or not. Along with label feature for each customers are given. Based on this data our model should learn pattern to identify defaulters so that if we give those features of any new customer or unseen customer, from which model was made, then model should classify where there is chance of customer to be a defaulter or not. This will help bank(or any other business related to this) to take action accordingly and to prevent loss which occurs because of credit defaulters.

3 Literature Review

Credit defaulters is one of main problem which bank face as not recovering money lends them to loss. If banks get to know probable defaulters then according they can take action and it will help them to prevent loss due to this.

In problem like this its almost impossible to get data where both label defaulter and non defaulter are almost same in number instead of this defaulters are very less as in our case its 6% of total data. So, data balancing is one of the important technique to include. Data balancing helps us in making model which not skewed to any particular class which leads to miss-interpretation.

4 Data

Variable Name	Description	Type
SeriousDlqin2yrs	**Person experienced 90 days past due delinquency or worse**	**Y/N**
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
DebtRatio	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

5 Deliverable

Model which will classify credit defaulter correctly.

6 Evaluation

Evaluation metric and validation techniques:
1. cross validation
2. Accuracy
3. F1 score
4. aur_roc

7 Data Ingestion

We are reading data from Data folder using pandas library as pandas data frame

8 Data Analysis

Some field are empty leading to missing value, so missing value treatment is required.
As this data is highly imbalanced we need to do data balancing.
In this data outlier detection and outlier treatment is also required as some values are unlikely e.g. age being 109 is very unlikely as present in this data.
Because of skewness of data, standardization is required.

9 Data Munging

As data is cleaned so this step is not required.

10 Data Exploration

Technique Applied:

1. Missing value treatment :
   As all are continuous so mean value is used to fill missing values.
2. Outlier treatment:
   Considering normal distribution in view outlier treatment is done using mean and standard deviation.
3. Data Balancing:
   "imblearn" library is used to do oversampling for minority class and under sampling for majority classs.

11 Feature Engineering

Following Data Exploration, mention feature engineering techniques. This includes feature selection.

12 Modeling

Python's 'scikit-learn' library is used for modeling.

Different models are used to compare there performance, these are:
1. logistic regression
2. svm
3. mlp
4. GBM

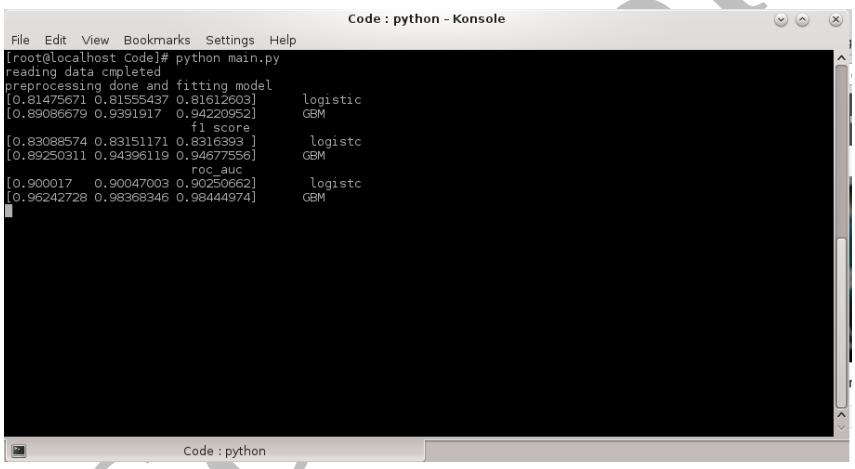
13 Optimization

Random search method is applied.

14 Prediction

Without random search best accuracy was by GBM: 94%
with random search best accuracy was by GBM: 96%

15 Visual Analysis



16 Results

[0.81257225 0.81140048 0.81460271] logistic
[0.92870145 0.92556471 0.92541855] svm
[0.71529743 0.54807335 0.63263354] mlp
[0.89119199 0.93801488 0.94214722] GBM
f1 score
[0.82924172 0.82745375 0.83038135] logistc
[0.9347059 0.93159617 0.9314811] svm
[0.75437096 0.84032258 0.74350274] mlp
[0.89266381 0.94305211 0.94655833] GBM
roc_auc
[0.89804437 0.89856208 0.9014922] logistc

[0.96799027 0.96672465 0.96732501] svm
[0.89877777 0.59556701 0.8660613] mlp
[0.96282017 0.98363341 0.98440272] GBM

logistic 0.832826641916
logistic {'penalty': 'l1', 'C': 30.10127353405051, 'fit_intercept': True}

GBM 0.952003507078
GBM {'max_features': 'auto', 'learning_rate': 0.3945925710914654, 'max_depth': 8}

f1 score
logistic 0.832826641916
logistic {'penalty': 'l1', 'C': 46.226359596839714, 'fit_intercept': True}

GBM 0.962003507419
GBM {'max_features': 'auto', 'learning_rate': 0.3945925710914654, 'max_depth': 8}

From result we can see that best model was given by GBM although SVM result was lose to GBM but as SVM took more time then GBM and gave good score and GBM performed best with random search too.