Loan Default Prediction

Capstone Project Report – Milestone 1 | Author: Shouvik Nandy

Contents

Problem Definition

Context	Home loans form major portion of profits for retails banks. Such loans when defaulted, result in big loses for the banks. Therefore it is significant that banks are diligent while approving the loans. Due to such high risks in the lending business, banks have traditionally employed strict processes to evaluate and approve loan applications. Such procedures are prone to flaws in human judgement and biases, they are also resource and time intensive. With the advent of data and machine learning there has been a shift to employ machine learning models to automate the approval of loans.
Objective	Develop a predictive model to simplify the loan approval process of a bank's consumer credit department. The model will be developed with data that the bank have obtained via its existing loan under writing process. The model must be human interpretable and provide justification for loan approvals and/or rejections.
Key Questions	In the process of model development , we will evaluate hypothesis and examine facts to a. Identify patterns in data that are important for prediction of loans b. Efficiency of the model
Problem Formulation	Develop a classification model to predict whether a customer will default on their loan, provide recommendations to the department of credit in the bank about the important features that has to be considered while approving new loans.

Data Exploration

Data Description

The Home Equity (HMEQ) dataset contains loan information recent home equity loans. The target variable is '*BAD*', its is a binary variable indicating whether a customer has defaulted on the loan.

File Name: Data Set.csv

File Type: Comma Separated Values

Record Count: 5960

Nbr of Target Variable: 1

Nbr of Independent Variables: 12

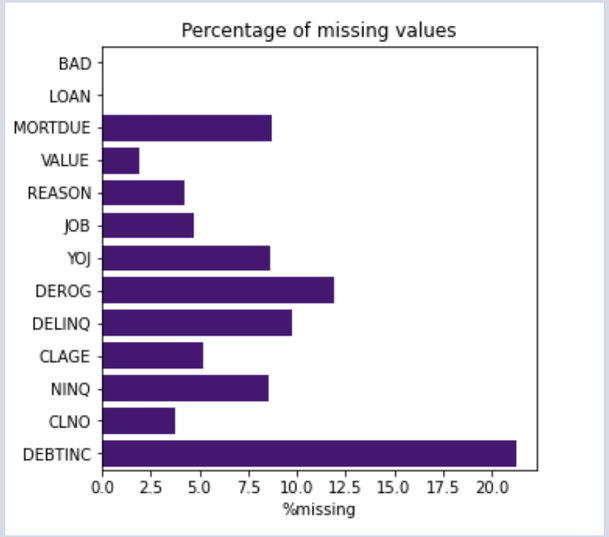
Imbalanced Dataset: Yes

Variable Name	Data Definition	Data Type*	Target Variable
BAD	Target variable for the model to predict. 1 = Client defaulted on loan 0 = loan repaid	Numerical	Yes
LOAN	Amount of loan approved.	Numerical	No
MORTDUE	Amount due on the existing mortgage.	Numerical	No
VALUE	Current value of the property.	Numerical	No
REASON	Reason for the loan request. HomeImp = home improvement DebtCon = debt consolidation which means taking out a new loan to pay off other liabilities and consumer debts)	String	No
JOB	The type of job that loan applicant has such as manager, self, etc.	String	No
YOJ	Years at present job.	Numerical	No
DEROG	Number of major derogatory reports (which indicates a serious delinquency or late payments).	Numerical	No
DELINQ	Number of delinquent credit lines (a line of credit becomes delinquent when a borrower does not make the minimum required payments 30 to 60 days past the day on which the payments were due).	Numerical	No
CLAGE	Age of the oldest credit line in months.	Numerical	No
NINQ	Number of recent credit inquiries.	Numerical	No
CLNO	Number of existing credit lines.	Numerical	No
DEBTINC	Debt-to-income ratio (all your monthly debt payments divided by your gross monthly income. This number is one way lenders measure your ability to manage the monthly payments to repay the money you plan to borrow.	Numerical	No

Data Exploration

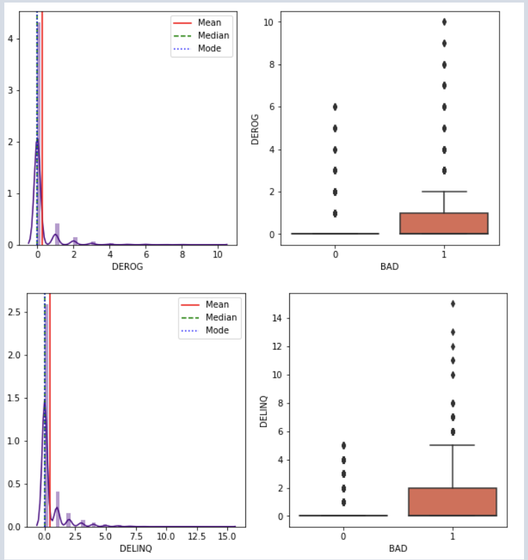
Observation & Insights

#1 Data Quality



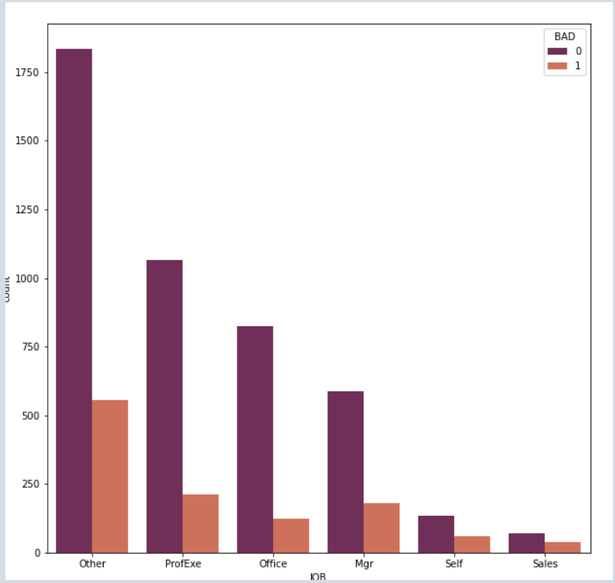
The missing values has to be managed by appropriate imputation techniques during data preparation.

#2 Distribution of DEROG & DELINQ variables



Customer who default loans have more derogatory reports and higher number of delinquent credit lines.

#3 JOB Categories



Customers with job category ‘Self’ & ‘Sales’ have higher percentage of defaulted loans

Proposed Approach

Potential Techniques

Clustering

Explore hidden patterns in data using PCA and t-SNE. Explore whether dimensionality reduction results in better performance of classification.

Baseline Model

Logistic regression & KNN model will be used as a baseline model for the prediction problem. Feature selection methods and hyperparameter tuning technique will be utilized to improve model performance. SMOTE technique will be utilized to handle imbalanced data for model training.

Decision Trees

Tree based model will be developed to explore potential improvement over baseline model. Feature selection methods and hyperparameter tuning technique will be utilized to improve model performance. Since decision trees have tendency of overfitting, we will explore generalization techniques of k-fold validation.

Ensemble Learning

Explore whether Boosting and Bagging techniques improve model performance. ADABoost and Random Forest Classifiers will explore to evaluate model performance improvement over Decision Trees and baseline model.

Proposed Approach

Solution Design

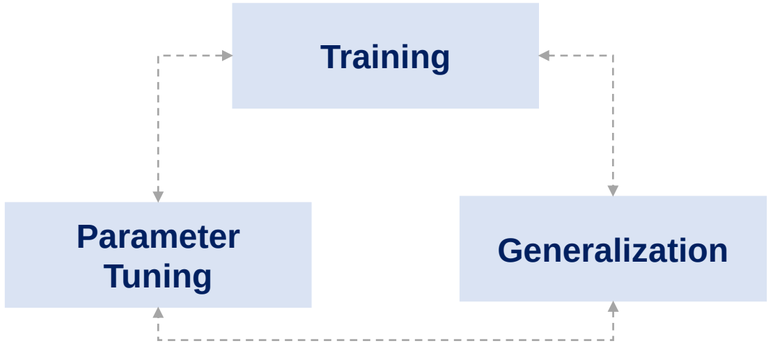
- *Handle missing values*
- *Encode Categorical variables*
- *Scaling of numerical variables*
- *Split data in train , validate and test sets*
- *Use SMOTE for oversampling of imbalance class*

- *Models: Logistic, Decision Tree, ADABoost & Random Forest*
- *Feature Selection*
- *Tune hyperparameters using Random and Grid methods*
- *Generalization & Regularization for overfitting using cross validation methods*
- *Utilized model performance metrics (AOC & Precision) for optimization*

- *Predict target class using the best model*

- *Measure Model performance*
- *Identify and report import features of the model*
- *Utilized SHAP for tree based models*

EDA › **Data Preparation** › **Model Development** › **Prediction** › **Model Evaluation**



- *Completed*
- *In-Progress*
- *Not Started*

Proposed Approach

Measure Of Success

Precision & Accuracy

Precision along with Accuracy will be used to evaluate model performance. Since we are dealing with a classification problem where the impact of incorrectly approving of a loan is critical than incorrectly rejecting the loan we will use Precision for model validation along with Accuracy.

AUC

We will use AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve to determine how well the model differentiates (separability) BAD loans.

Explainable

Since this problem deals with a financial service to a customer it is important to explain the reason behind adverse results of the model. This will include clear explanation of the important features that are important for model prediction.

Appendix

Notebook	Description	Format
Capstone_Loans_Default_Prediction_Shouvik.ipynb	Notebook with EDA related to Milestone 1 deliverables.	IPYNB
Capstone_Loans_Default_Prediction_Shouvik.html	Notebook with EDA related to Milestone 1 deliverables.	HTML