Smart Banking analysis

1 Technical Design Document

Version1.0

Contributors

The Content of this document has been authored with the combined input of following group of key individuals.

|  |  |
| --- | --- |
| **Name** | **Section Worked Upon** |
| Rajesh Venaganti | Initial Draft |
| Mahesh | Initial Draft |
|  |  |

1.Introduction

The goal here is to build an end to end automated Machine Learning solution where the user will only give the data and select the data option for single or bulk prediction, and the result will be the prediction whether a patient is Diabetic or not based on the model. This is a Regression problem and we will be building our model based on a Regression Algorithm

The technical design document gives a design blueprint of the Early Stage Loan Elgibility Prediction project. This document communicates the technical details of the solution proposed.

In addition, this document also captures the different workflows involved to build the solution, exceptions in the workflows and any assumptions that have been considered.

Once agreed as the basis for the building of the project, the flowchart and assumptions will be used as a platform from which the solution will be designed.

Changes to this business process may constitute a request for change and will be subject to the agreed agility program change procedures.

**Note: All the code will be written in python version 3.6.9**

2.Dataset Information

|  |  |
| --- | --- |
| Attribute ID | Values |
| Loan ID | Loan id |
| Customer ID | Customer ID |
| Loan Status | Fully Paid, Charged off |
| Current Loan Amount | 10802-99999999 |
| Term | Short term, Long term |
| Credit Score | 585 to 7510 |
| Annual Income | 76627-165557393 |
| Years in current job | <1 year to 10 years + |
| Home Ownership | Home Mortgage, Own home, Rent |
| Purpose | Home Improvements, Debt Consolidation, Buy House, Other, Business Loan, Buy a Car, major purchase, vacation, small business, Medical Bills, wedding, Educational Expenses, moving, renewable energy |
| Monthly Debt | 0 to 107025.67 |
| Years of Credit History | 3.6 to 70.5 |
| Months since last delinquent | 0 to 176 |
| Number of Open Accounts | 0 to 76 |
| Number of Credit Problems | 0 to 15 |
| Current Credit Balance | 0 to 5850651 |
| Maximum Open Credit | 0 to 489343206 |
| Bankruptcies | 0 to 7 |
| Tax Liens | 0 to 15 |

3.Dataset variables Description

**Loan id:** Loan id evaluates identification of particular loan . it’s doesn’t impact on target variable

**Customer id**: Customer id evaluates identification of particular loan . it’s doesn’t impact on target variable.

**Loan Status:** This feature analysis of whether customer loan paid fully or not. Loan charged off explains the that particular customer is a bad debt person(Uncollectable loans)

**Current Loan Amount:** Based on this feature we can analyses eligibility of Loan amount of particular customer, If suppose customer loan creditability with company is high, then that particular eligibility to take less loan amount only, By comparing correlation with Annual income feature we can offer the loan, To offer the loan Current loan amount feature should be in inadequate to Annual income.

**Term:** if loan maturity is Long term we can analysis as very high obligation for company to recover the loan, If loan maturity is short term we can analysis as very low obligation for company to recover debt from customer

**Credit Score:** If customer have high credit score in less transaction history , then we can estimates we have less liability to recover the loans with that customer, so we can offer the loan.

**Annual Income:** Based on this feature we can estimates liability of loan repayment of the customer.

**Years in Current job:** This feature vales estimates stability of customer income. If customer have higher experience, then it shows financial stability of customer.

**Home Ownership**: This feature have three variables i.e Home Mortgage, Rent, Own home. If customers have Home Mortgage, Rental expenses, then they have high obligation to repay loan compare to own home customers. These variable explains liability of repayment of loans.

**Purpose:** This feature variables explains loan investment of customer. If suppose customer investing his loan into any investments we can expects returns from investments. So there is less obligation to the customer to repay the loan. So it’s very important to analyse the loan purpose of the customer.

Monthly debt: If customer have less amount of debt then it’s less obligation to customer to repay the loan.

**Years of Credit history:** If customer have good credit history, then companies have less obligation to recover loans, So offering loan to the customer doesn’t seems to be obligation

**Monthly of delinquent:** In the personal finance field, the term "delinquent" commonly refers to a situation where a borrower is late or overdue on a payment, such as income taxes, a mortgage, an automobile loan, or a credit card account. People who are late with a credit card payment may be forced to pay a late fee.

**Number of Open Accounts:**

**Number of Credit Problems:** If Number of credit problems count less, then company have less obligation to recover loan, So it’s have highly correlated with other variables

Current Credit balance: High Credit balance of customer of evaluates the less obligation to recover the loans to the company.

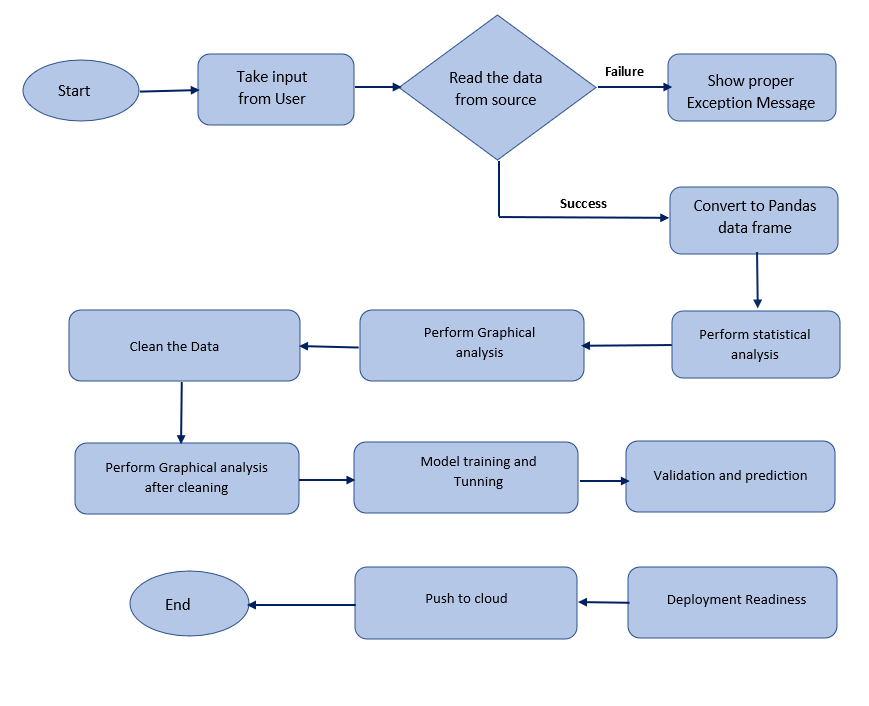
**Bankruptcies:** Bankruptcy is a legal proceeding involving a person or business that is unable to repay their outstanding debts. The bankruptcy process begins with a petition filed by the debtor, which is most common, or on behalf of creditors, which is less common. All of the debtor's assets are measured and evaluated, and the assets may be used to repay a portion of outstanding debt.

**Tax liens**: A tax lien is a legal claim against the assets of an individual or business that fails to pay taxes owed to the government. In general, a lien serves to guarantee payment of a debt such as a loan, or in this case, taxes. If the obligation is not satisfied, the creditor may proceed to seize the assets.

4.High Level Objectives

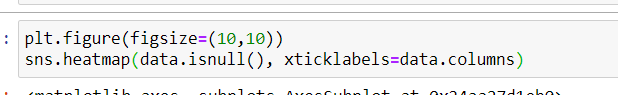
1. Enable reading/loading of data from the source and convert them into pandas dataframe.
2. Importing all the relevant libraries to perform Data Analysis Task.
3. Perform statistical analytics of the data and prepare a table for the analysis and show it on screen.
4. Performing graphical analysis of the data.
5. Handling Missing Values of the data.
6. Find Outliers and removal of the same using appropriate techniques.
7. After data cleaning showcase the graphical analysis once again for comparison.
8. Data transformation for further Analysis i.e. categorical to binary, label encoding, One hot encoding etc.
9. Standardizing the numerical fields into its Standard Normal Form.
10. Splitting the Data for Training and Testing purpose.
11. Importing pipeline for comparison between many ML models.
12. Choose the appropriate ML model for training based on classification report , accuracy and level of significance between training and testing data.
13. Tuning our model by Hyper Parameter Optimization.
14. Saving our model for prediction.
15. Create a Flask based Web application for the user and give options for single prediction as well as bulk Prediction to perform the tasks of:
16. Validation of the data provided by the user.
17. Pre-processing of the data provided by the user converting into appropriate encoded form and also standardization of the same
18. Prediction by the model on the given data
19. Deployment of the Flask based Web application for prediction in an appropriate Cloud Platform.
20. Creation of Docker container for the application

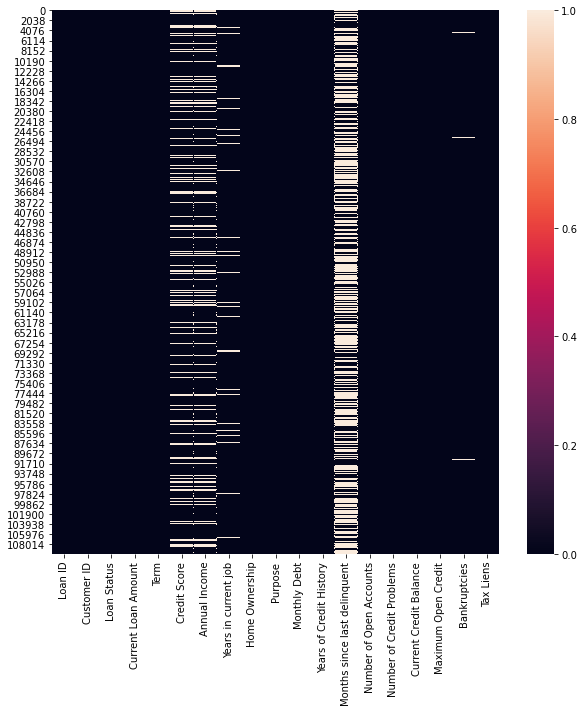
5.Workfow overall



6.EDA

1. Finding the information about attributes, There are 19 columns out of Categorical variables which are Loan ID, Customer ID, Loan status, Term, Years in current job, Home Ownership, Purpose and remaining 12 columns are Numerical columns.
2. Finding the missing values: Here we used Graphical representation to find the missing values in columns using Heatmap. Here we can observe Credit score, Annual Income, Years in current job, Months since last delinquent, Bankruptcies, Tax liens have null values

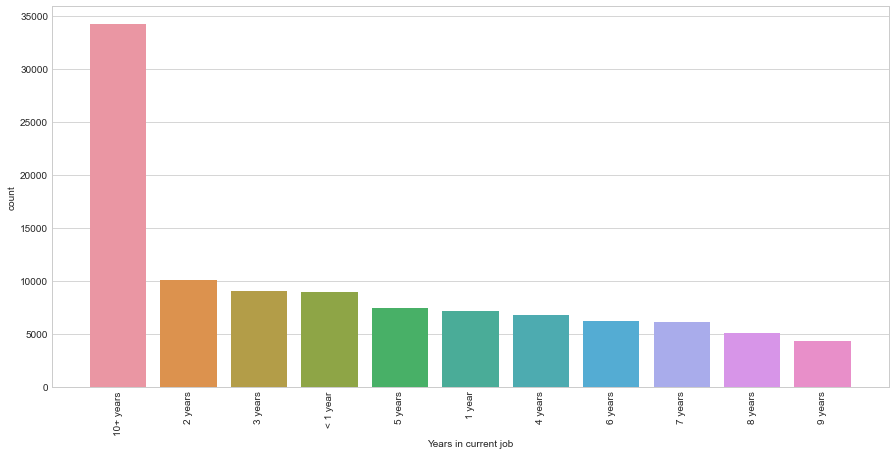




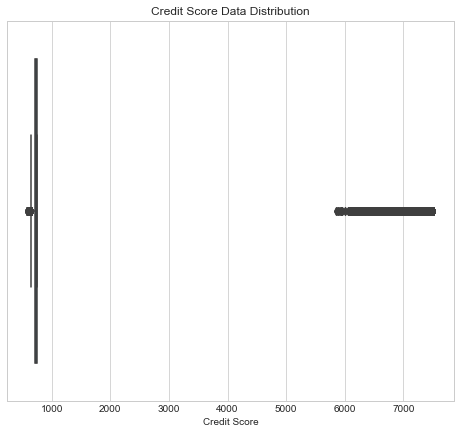
1. From above data Loan id , Customer id doesn’t impact on target variables so we are dropping the these columns from data.
2. Assuming blank values of nan as zero delinquency as of now Before that saving the current

data file if we want to try later different approach for replacing missing values.

1. From below data shows majority of the data has a work experience of 10 + years so we can replace missing experiences with 10 + years. Also as of now work experience is in character format , we will replace them with the numeric format which will ease our analysis .



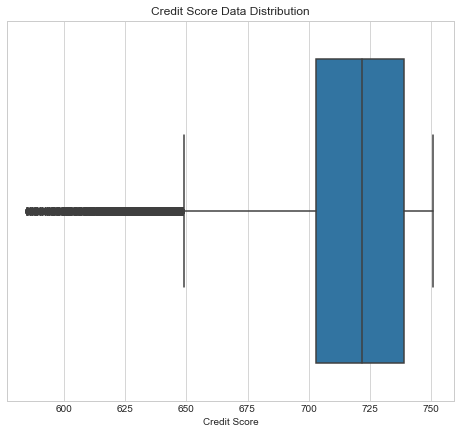
1. Credit Score data distribution: From the description as well as from the box plot it is clear that most of the credit score data varies around credit score of 700 .



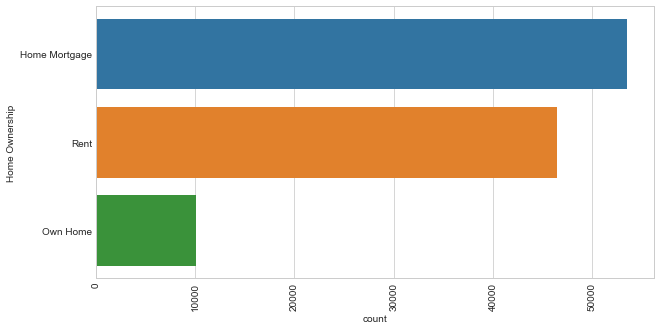
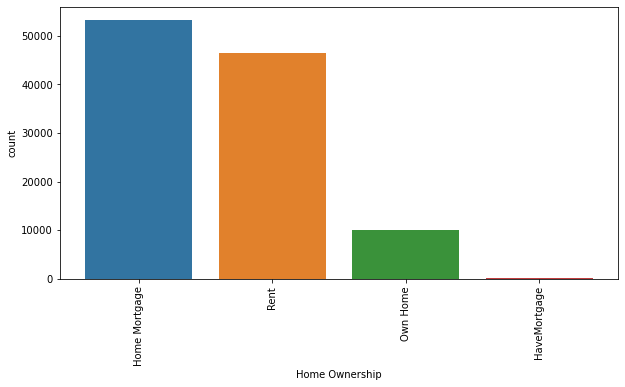
Above shows around 5000 records which have values beyond credit score of 900 which looks like a typo error. Also apart from this we have around 20k+ missing values which also needs to be addressed but problem is if we address the missing values without handling the

incorrect credit scores then it might impute missing values with incorrect values. Reason is the incorrect (typo error) credit scores are big numbers and they will influence the calculations with higher weightage.

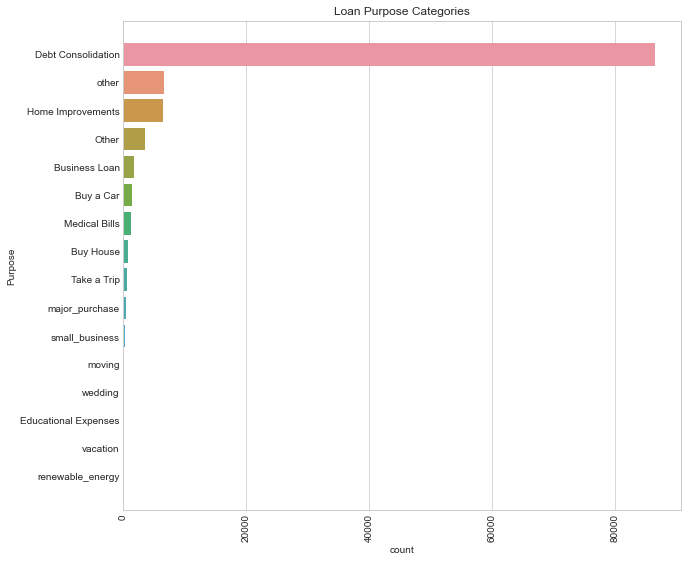
Below we can see the extreme values are handled now. With values varying from 500 to 750.



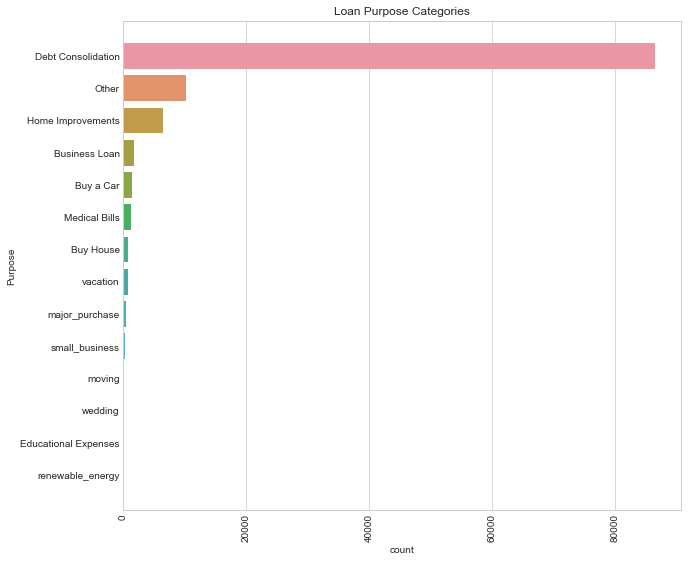
1. Distribution of Home Ownership: In below image there are four unique values in Homeownership columns such as Home Mortgage, Rent, Own Home, Have Mortgage, Here Looks like there is a type for Home Mortgage which is typed is Have Mortgage, So we are replacing the values Have Mortgage with Home mortgage



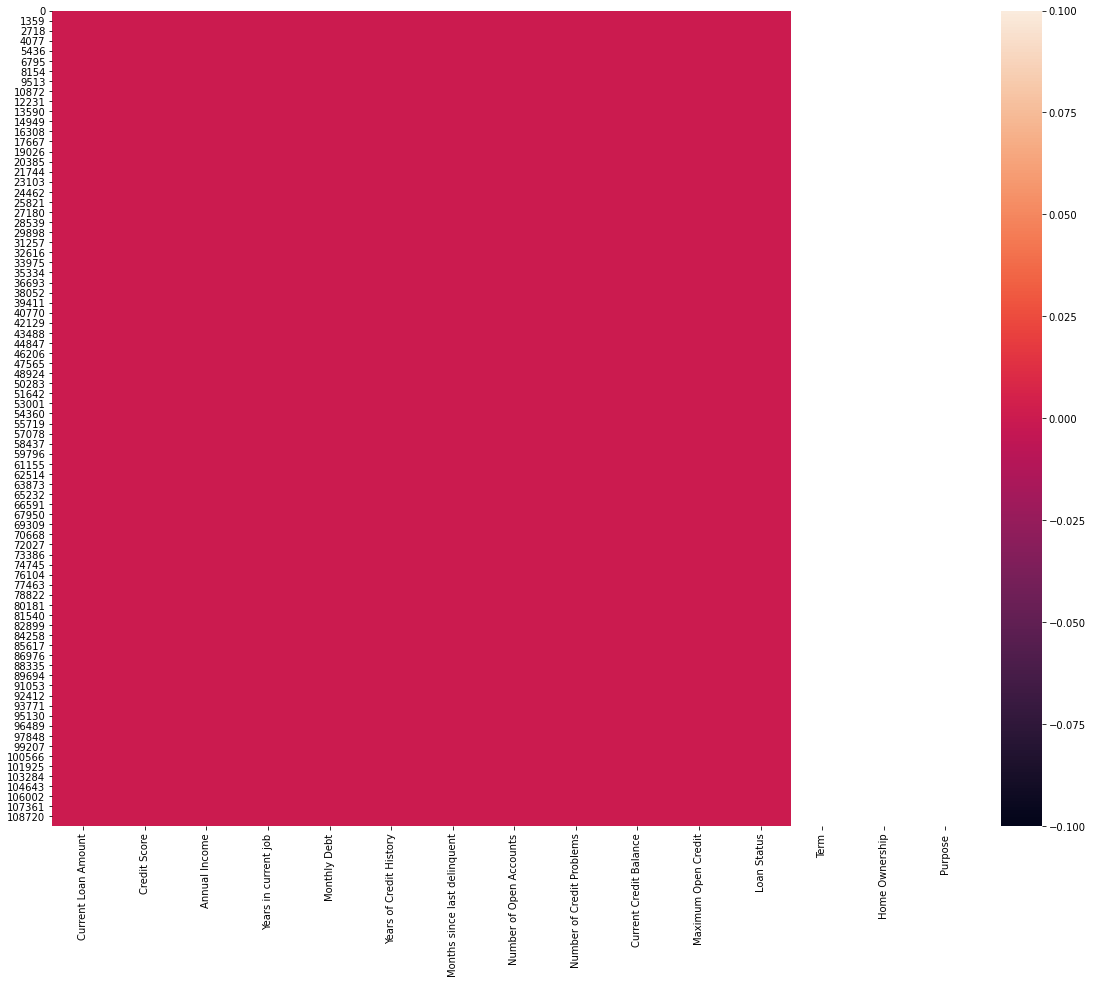
1. Distribution of Purpose column and Manipulation: Below we can see there are some categories which are either written twice or very similar which can be combined .This will also help in reducing the number of features which will result when we do encoding to handle categorical features.







1. To replacing missing values in Data frame we are using KNN imputer



7)Model Selection and Tuning

1. Splitting the dataset into training and testing dataset.
2. Use of Pipeline for comparisons among the different classification algorithms –Linear Regression, Decision Tree Classifier, XGBoost based on accuracy , classification report and level of significance among training and testing data.
3. Here we have chosen XGBoost because it is based on ensemble approach based on Gradient Boosting where each new tree helps to correct errors made by previously trained tree and gives scalability and robust way of prediction.
4. We have used hyperparameter optimization for model tuning – Here we have used a combination of RandomizedSearchCV , GridSearchCV and TPOT which is a genetic search algorithm based to find the best parameters of tuning.
5. After finalizing the model we have saved the model in a pickle file.

