**COMSATS University Islamabad,   
Park Road, Chak Shahzad, Islamabad Pakistan**

Project Report

for

**Bank Loan Prediction**  
Version 1.0

**By**

**Syed Ali Gilani CIIT/SP17-BSE-079/ISB**

**SupervisorMr. Qasim Malik**

*Bachelor of Science in Software Engineering (2017-2021)*

|  |  |  |
| --- | --- | --- |
| **No.** | **Comment** | **Action** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**SCOPE DOCUMENT REVSION HISTORY**

**Supervisor Signature**

**Date:**

**Table of Contents**

Abstract 3

1. Introduction: 4

2. Problem Statement: 4

3. Solution by ‘Bank Loan Prediction’ 4

4. Advantages/Benefits of MarkoVoice 4

5. Scope 4

6. Modules 4

7. Dataset 5

8. System Limitations 5

9. Conclusion 5

**Project Category:**

* Problem Solving and Artificial intelligence
* Machine Learning

# Abstract

‘Bank Loan Detection’ is an application which will predict whether or not a bank should give loan to its customers. It will predict the loan is risky or safe.

# Introduction:

‘Bank Load Prediction’ is an application that will apply classification, logistic regression and decision tree to implement the bank load prediction system. This application will predict whether a loan is risky or safe. The Banks give loans to their customers and sometimes they have to face difficulty on recovering the amount given. Sometimes banks bear loss when their customers flee without paying back the bank loan. The application will help banks avoid such situation and use real world datasets to train a model that can predict whether loan is risky or safe.

data set is taken from lending club. Lending Club is the platform or rather the marketplace where investors and borrowers meet virtually. Lending Club processes the application with their own data science methods but on the side of the investor there is still due diligence to be performed in order to insure the creditworthiness of the borrower and the level of risk involved in any given case. Applying machine learning to loan default predictions showcase a useful application of this branch of artificial intelligence to solve real-world and business problems. We will try to build this model with the most transparency possible as to mirror the conditions in which financial institution must disclose this process. Even though we are building this model with a conservative investor in mind, the model could be later tweaked as to be used in a financial institution or a Lending Club investor

# Problem Statement:

Giving loan to public, corporate and private sectors is an essential part of daily banking. The Banks give loans to their customers and sometimes they have to face difficulty on recovering the amount given. Sometimes banks bear loss when their customers flee without paying back the bank loan. Therefore, it is very necessary for banks to analyze whether they should give loans to their client or not. And what will be the probability of returning loan by clients.

# Solution by ‘Bank Loan Prediction’

Application apply some logistic regression and classification algorithms of machine learning to create a solution for prediction of loans. The application will train a model on real dataset that will be able to predict whether a loan is safe or risky.

# Advantages/Benefits of MarkoVoice

Following will be the key advantage of the Bank Loan Prediction (Theoretically mentioning):

* Banks will know in advance whether the loan given is safe or risky
* Banks can check the probability of loan returning by clients
* Banks can maximize their cash flow and profits by giving loans to appropriate parties.

# Scope

‘Bank Loan Prediction’ application will train a model on real dataset. The dataset will be taken from kaggle.com which is public website for providing datasets. Some logistic regression plus classification algorithm will be applied. Decision tree will also be used if needed depending on the progress of project. A model will be trained by above mentioned algorithms to predict loan. After Training some un-known data will be given to application to get the prediction. Dataset used will not be in raw form.

# Modules

**Module1: Data Preprocessing**

data set is in raw form so it was necessary to perform data cleaning in order to improve the quality of dataset. Some step involved in data pre processing are:

1. Exploratory data analysis
2. Data Cleaning
3. Selecting relevant features
4. Null value imputation
5. Creating dummy variables
6. Handling outliers (partially working)
7. Multi-co-linearity Check (partially working)

**NOTE**: over all this module is working fine

**Module1: Model Training:**In this module the application will be trained via classification algorithms. Datasets for the training of this module will be taken from kaggle.com. In this module algorithms will be applied on dataset and training will be perform to make classes of ‘safe’ and ‘risky’. Loan will be either ‘safe’ or ‘risky’. Four steps involved are:

1. Data preprocessing
2. Training ML Model (partially working)
3. Evaluation model on test data (Not working)
4. Learning curve and scope for improvement (Not working)

**NOTE**: This module is not working 100% only partially working. I was unable to complete the task due to heavy data set and high algorithm complexity

**Module3: Application integration with backend(GUI)**

Application to provide the interface was to be done in anvil but due to partial working of the module 2, it is useless to make frontend. Backend ran into some errors which I cannot resolve and I cannot complete the backend so I did not created the application front end.

# Dataset

Dataset is public which is available on lendingClub.com and also on kaggle.com. I have downloaded dataset and uploaded to my drive whose link is shared below.

Links:

Dataset 1: <https://drive.google.com/file/d/1B6sFf60o_Qe46fIhvAmC8vWDieVfgmCJ/view?usp=sharing>

Dataset 2: <https://drive.google.com/file/d/1f47YbI3eopFEtyQ5Hr6sjFFLu6vOwZCD/view?usp=sharing>

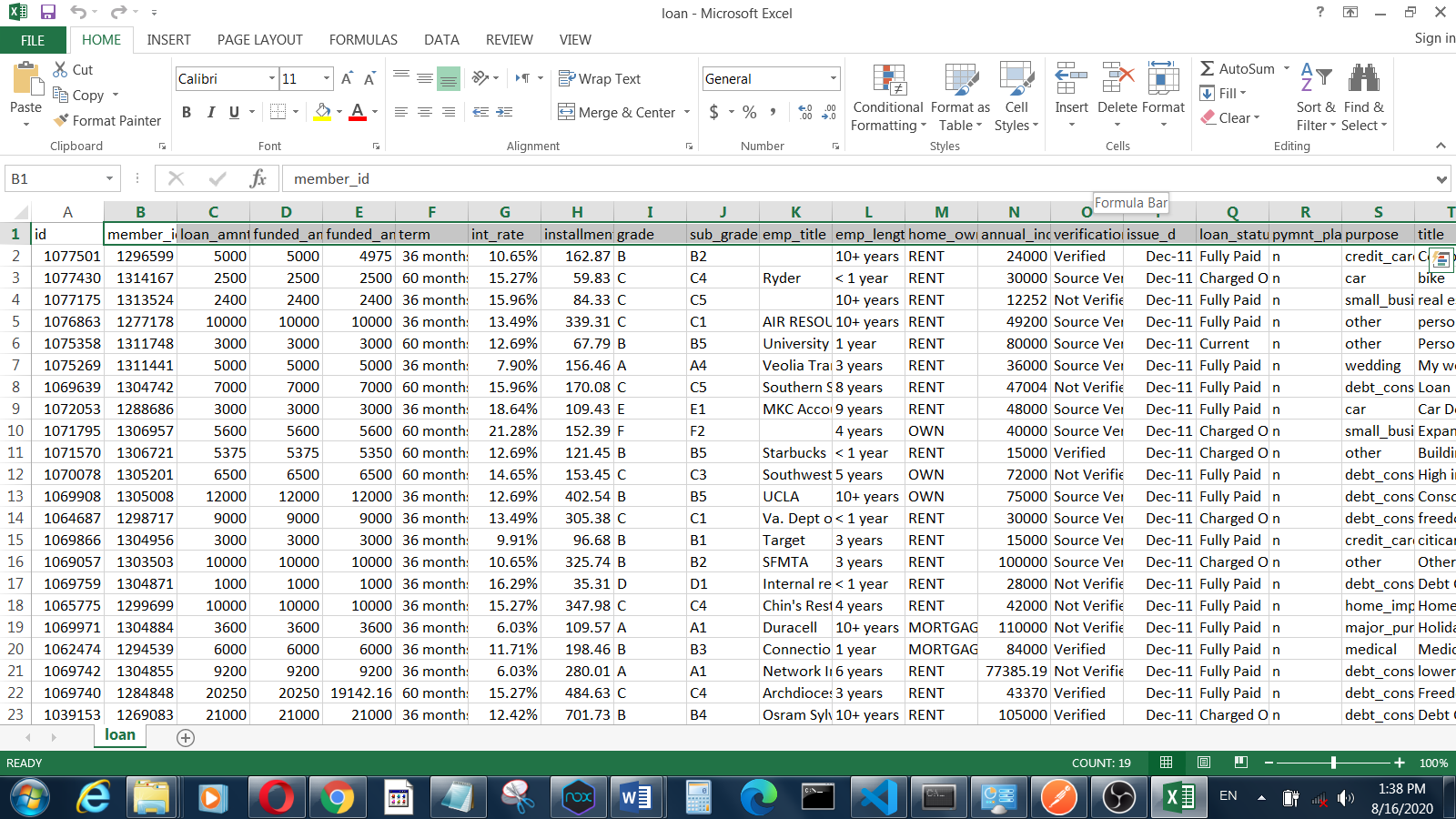
Data Set screenshots:

Figure 1: loan\_data.csv

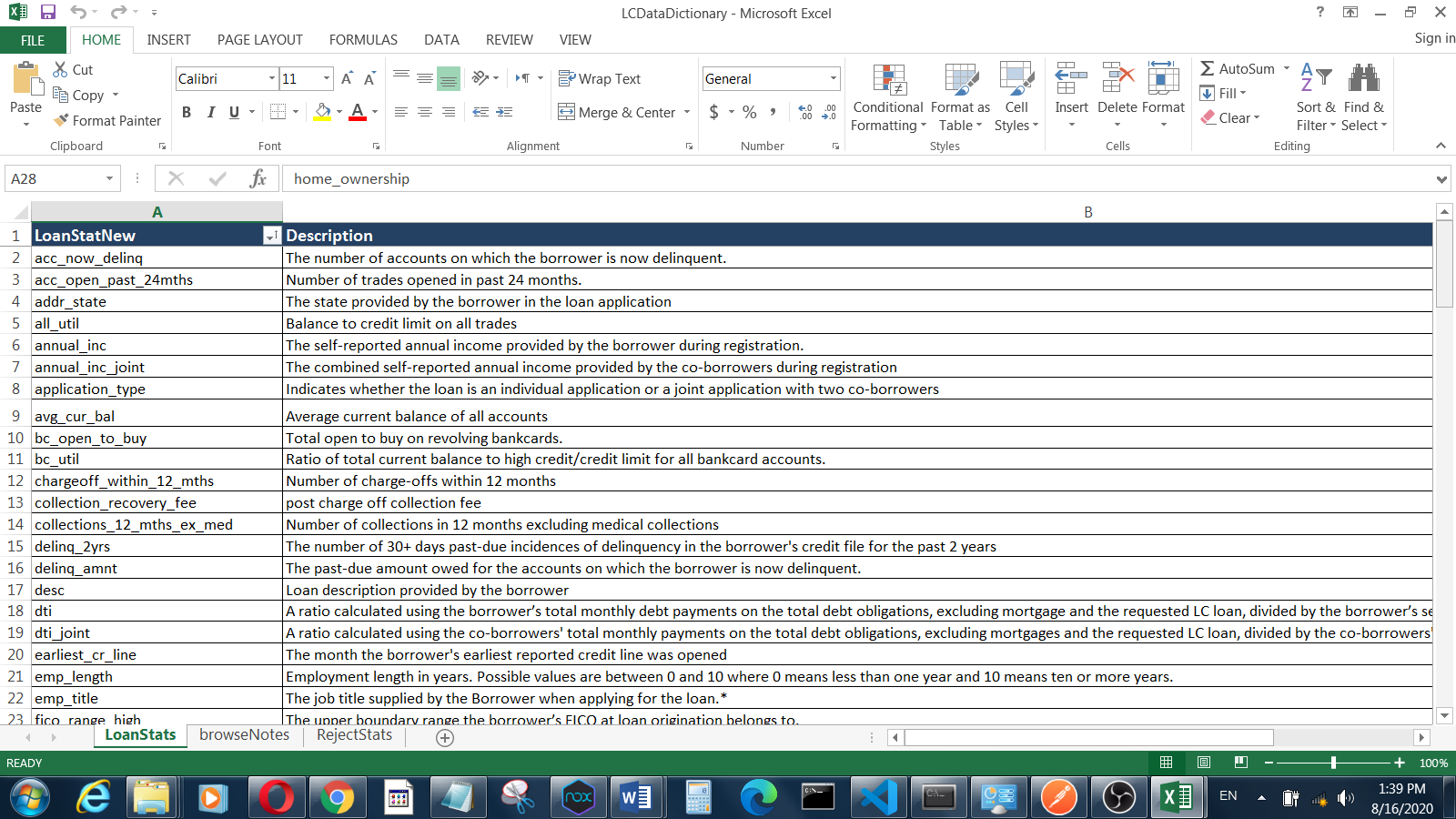


Figure 2: LCData\_Dictionary.xlsx

# Tools & Technology

**Google colab** was used for development. Data set was referenced from drive and all the processing was being done on google colab. Some custom classes and functions were also used from lending club.

**Table 1: Tools and Technologies**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tools**  **And**  **Technologies** | **Tools** | **Version** | **Rationale** |
| Google Colab | - | Online IDE; easy to run |
| MS Word | 2019 | Project Report |
| **Technology** | **Version** | **Rationale** |
| Python | 3.7 | Programming language |
|

# Implementation

# First of all data cleaning was done. Then model training and evaluation and finally prediction and testing model.

# Data cleaning

1. Model Training
2. Data preprocessing
3. Training ML Model (partially working)
4. Evaluating Model and Test Data (Not included)
5. Learning curve and scope for implementation (Not included)

# Data cleaning was done in 4 to 5 steps. These steps are selecting relevant features, null value imputation, creating dummy variable, handling outliers etc. This process worked fine. Then comes the Model training and evaluation.

NOTE: model training and evaluation is not working properly due to heavy data set and complexity of implementation. I was unable to finish the project.

# Source Code

Source code available on GitHub.  
link: <https://github.com/makhdoomAliGilani/Bank-Loan-Prediction>

# Web Application of Project

Application was not developed because backend was not finished and model was not properly trained due to some complexities. I was unable to finish the backend so it was useless to make application front end.

# Conclusion

‘Bank Loan Prediction’ is a project which will be very beneficial in learning machine learning algorithms through the view point of implementation, reflecting real world problems..