Georgia State University

J. Mack Robinson College of Business



Bank Personal Loan Prediction

Spring 2022

**By:**

**Sai Charan Pappala**

**Sukumar Reddy Kottala  
Narender Reddy Konuganti**

**Gopi Sandeep Guntuku**

Table of Contents

[1. Project Description 3](#_Toc84186434)

[2. Goal & Business Problem 3](#_Toc84186435)

[3. Data Exploration & Preprocessing 3](#_Toc84186436)

[4. Data Cleaning](#_Toc84186437) 3

[5. Graphs and plots to explore the Data](#_Toc84186438) 4

[6. Reasons for specific Model Selection 9](#_Toc84186440)

[7. Results](#_Toc84186441) 10

8. Summary………………………………………………………...…………….10

**Project Description:**

This project is about predicting likeliness of converting liability customers to personal loan customers using Logistic Regression, Random forest, KNN, Neural Networks, and Ensemble Models using “**Bank Personal Loan Modelling**” Dataset from Kaggle (<https://www.kaggle.com/krantiswalke/bank-personal-loan-modelling>) based on 5000 observations with 14 explanatory variables.

**Goal:**

The project is aimed at implementing a model(s) to predict likeliness of converting liability customers to personal loan customers.

* Remove variables, build what is needed.
* Models: Logistic Regression, KNN techniques, RandomForest, Ensemble Learning & Neural Networks.
* Choose the best model having best accuracy.

**Business Problem:**

This case is about a bank whose management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

**Data Exploration and Preprocessing:**

* The Dataset contains data of 5000 customers with 14 explanatory variables.
* The data includes:
* Customer demographic information (age, income, etc.),
* The customer's relationship with the bank (mortgage, securities account, etc.)
* Customer response to the last personal loan campaign (Personal Loan).
* Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

**Data Cleaning:**

* Removed rows which were having unknown values for features like Zip code and ID.
* Dropped rows with Nan/Null values.
* Dropped index column
* Checking for outliers, data entry errors
* Apply abs for “Experience”

**Graphs and plots to explore the data:**

1. **HISTOGRAM OF ALL FEATURES:**

Chart, histogram

Description automatically generated

1. **BOX PLOT:**

Chart, box and whisker chart

Description automatically generated

1. **DENSITY PLOT:**

Graphical user interface, application

Description automatically generated

1. **HISTOGRAM OF LOAN DATA SAMPLING:**

Chart, waterfall chart

Description automatically generated

**Models and their comparison:**

We have implemented the below models:

* Logistic Regression
* RandomForest
* Classification using K-Nearest Neighbors
* Neural Networks
* Ensemble method

## Reasons for specific Model Selection:

## Logistic Regression:

Since we are dealing with a classification problem and expect some linear relationships

between variables, we will use a logistic regression model to classify our data. The Logistic Regression model on the testing data gives an **accuracy** value of **90.6%.**

Graphical user interface, text, application

Description automatically generated

**Classification using K-Nearest Neighbors:**

KNN stands for **K-Nearest Neighbors**. It is a supervised learning algorithm. It is often used as a benchmark for more complex classifiers such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM). We have used 14 independent features for KNN implementation. A robust implementation must consider feature engineering, data cleaning, and cross-validation.

* **K means clustering**
* K = 3
* Sampling 80% of data for training the algorithms using random sampling

We have implemented KNN with different optimal weights by changing k values and this time the **accuracy** we achieved is **99.2%.**

Chart, line chart

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

**RANDOM FOREST:**

* RandomForest
* Type of randomforest: classification
* Number of trees: 500
* No. of variables tried at each split: 3

Accuracy: 99.2%

**Timeline

Description automatically generated**

Table

Description automatically generated

Table

Description automatically generated

**Neural Network:**

Neural networks are a class of machine learning algorithms used for complex patterns in datasets using multiple hidden layers and non-linear activation functions. They are also known as artificial neural networks (ANNs) or simulated neural networks (SNNs). We have implemented in our scenario and the **accuracy** we achieved for the testing set is **91.2%.**

Text

Description automatically generated

**Ensemble: VOTING & WEIGHTED**

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. We have implemented ensemble techniques with three models: Logistic Regression, Neural Network and KNN in this project. The **accuracy** we attained 99.2% for both voted and weighted models but the sensitivity is 99.45 for weighted Ensemble.

Text

Description automatically generated

Text

Description automatically generated

## Results:

## Below is the accuracy for all the five models implemented in the project:

|  |  |
| --- | --- |
| **MODEL** | **ACCURRACY** |
| Logistic Regression | 90.6 % |
| KNN | 99.2% |
| RandomForest | 99.2 % |
| Neural Networks | 91.2 % |
| Ensemble | Voting: 99.2 %  Weighted: 99.2 % |

KNN and RandomForest performed the best with an accuracy of 99.2% followed by Ensemble and with an accuracy of 99.2%.

**Summary:**

The aim of the bank is to convert their liability customers into loan customers. They want to set up a new marketing campaign; hence, they need information about the connection between the variables given in the data. Four classification algorithms were used in this project. From the implementation, it seems like **KNN** have the highest accuracy and we can choose that as our final model.