Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Bank Prediction and Detection

CHITKARA UNIVERSITY

Supervised By:

Ms. Shagun Sharma

Submitted By:

Janvi Sethi (2210990431)

Jashanjit Kaur (2210990440)

Jasleen Kaur (2210990447)

Kashika (2210990493)

Department of Computer Science and Engineering
Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab

INDEX:

Abstract:

Machine learning is a field of artificial intelligence that involves training models to learn from data and make predictions or decisions without being explicitly programmed. There are different types of machine learning, including supervised learning, unsupervised learning, and reinforcement learning.

In this project, we implement machine learning techniques for four critical applications: loan approval prediction, loan status prediction, credit card fraud detection, and bank customer churn prediction.

For loan approval prediction, we utilize supervised learning, training models on labeled data that includes various attributes such as applicant income, credit history, loan amount, and employment status. The goal is to predict whether a loan application will be approved or rejected based on the learned patterns and relationships between these features and the approval status.

Credit card fraud detection focuses on identifying potentially fraudulent transactions. By training models on historical transaction data, including features like transaction amount, location, and time, the model learns to distinguish between legitimate and fraudulent activities. This enables the model to flag suspicious transactions in real-time, thereby helping to prevent financial losses.

The application of Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine models in these areas demonstrates the potential of machine learning to enhance decision-making, improve security, and increase customer satisfaction in the banking industry.

1. Logistic Regression:

- **Description:** Logistic Regression predicts binary outcomes, like loan approval or fraud detection, based on the relationship between dependent and independent variables.
- **Advantages:** Simple to implement and interpret, efficient for small datasets with limited features.
- **Disadvantages:** Can't capture complex relationships in data, prone to underfitting.

2. Decision Tree:

- **Description:** Decision Tree are non-linear models that recursively split the data based on features to create a tree-like structure. Each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label.
- **Advantages:** Easy to interpret and visualize, can handle both numerical and categorical data.
- **Disadvantages:** Prone to overfitting, sensitive to small variations in the data.

3. Support Vector Machines:

- **Description:** SVM is a powerful supervised learning algorithm that can be used for both regression and classification tasks. It finds the hyperplane that best separates the classes in the feature space, maximizing the margin between the classes.
- **Advantages:** Effective in high-dimensional spaces, memory-efficient, effective in cases where the number of dimensions is greater than the number of samples.
- **Disadvantages:** Not suitable for large datasets, computationally expensive, sensitive to the choice of the kernel parameters.

4. RandomForest:

- **Description:** Random Forest combines multiple decision trees to improve prediction accuracy for tasks like loan approval, fraud detection, and customer churn.
- **Advantages:** Random Forest provides high accuracy, robustness to overfitting, handles large datasets, and can model complex relationships in banking predictions.
- **Disadvantages:** Random forests can be computationally expensive, prone to overfitting with noisy data, and challenging to interpret due to complex interactions.

1. Introduction:

In the realm of banking, where financial decisions profoundly impact individuals and institutions, the integration of machine learning has revolutionized traditional practices. This project delves into a multifaceted exploration, focusing on crucial areas like loan approval prediction, loan status prediction, credit card fraud detection, and bank customer churn prediction.

With a comprehensive dataset encompassing diverse attributes such as applicant demographics, financial histories, transaction patterns, and customer behaviors, this project aims to harness the power of machine learning algorithms. The objective is to develop predictive models that can accurately assess the likelihood of loan approval, predict the status of existing loans, detect fraudulent activities in credit card transactions, and forecast customer churn within banking services..

Through meticulous data analysis, feature engineering, and model training, this project seeks to identify patterns, correlations, and predictive indicators that can significantly enhance decision-making processes in the banking sector. By evaluating and comparing various machine learning techniques such as Logistic Regression, Decision Trees, Support Vector Machines, and possibly others, the project endeavors to pinpoint the most effective models for each task.

The ultimate goal is to empower banking institutions with sophisticated tools that enable early risk detection, efficient resource allocation, and proactive customer engagement strategies. By leveraging machine learning insights, this project strives to optimize loan approvals, mitigate fraud risks, and enhance customer retention efforts, thereby fostering a more resilient and customer-centric banking ecosystem.

1.1 Background:

In today's dynamic financial landscape, the banking sector faces significant challenges in managing risk, ensuring customer satisfaction, and optimizing operational efficiency. Machine Learning has emerged as a transformative technology offering sophisticated solutions to these challenges. This project focuses on leveraging ML algorithms for four critical tasks in banking: loan approval, status, bank customer churn prediction and credit card fraud detection.

Loan approval prediction involves analyzing historical loan data to build models that assess the creditworthiness of applicants, aiding in making informed decisions on loan approvals. Loan status prediction extends this by predicting the likelihood of timely repayments, aiding in risk management and improving loan portfolio performance.

Credit card fraud detection utilizes ML algorithms to detect fraudulent transactions in real-time, enhancing security and minimizing financial losses. Bank customer churn prediction utilizes customer data to forecast the likelihood of customers switching to competitors, enabling proactive retention strategies and improving customer loyalty.

This project aims to showcase the efficacy of ML in addressing key banking challenges, ultimately contributing to enhanced decision-making, risk mitigation, and customer-centricity in the banking industry.

1.2 Objective:

The objectives of the project outlined in the provided code can be summarized as follows:

- **Loan Approval Prediction:** Develop a machine learning model to predict loan approval status based on applicant information**.** Evaluate the model's accuracy, precision, recall, and F1 score to assess its Performance. Analyze factors contributing to loan approval or rejection, such as credit score, income level, employment status, and loan amount.
- **Loan Status Prediction:** Build a predictive model to forecast loan status, including current, delinquent, or paid-off loans. Investigate key features influencing loan status changes over time. Implement strategies to improve loan status prediction accuracy and reduce false predictions.
- **Credit Card Fraud Detection:** Design and implement a fraud detection system using machine learning algorithms. Identify patterns and anomalies in credit card transactions indicative of fraudulent activities. Enhance fraud detection capabilities through feature engineering and algorithm optimization.
- **Bank Customer Churn Prediction:** Develop a churn prediction model to forecast the likelihood of customers leaving the bank**.** Explore customer behavior patterns, transaction histories, and engagement metrics to predict churn. Propose retention strategies based on predictive insights to mitigate customer attrition

- **Model Building:** We have used Support Vector Machines (SVM), Decision Trees, Random Forest, and Logistic Regression for our machine learning project related to banking, including loan approval prediction, credit card fraud detection, and bank customer churn prediction.

Evaluate the combined impact of loan approval prediction, loan status prediction, fraud detection, and churn prediction on the bank's operations and customer experience.

.

1.3 Significance:

The implementation of machine learning in these critical are offers several advantages:

- **Improved Decision-Making:** By leveraging machine learning models, banks can make more informed and data-driven decisions, leading to better risk management and resource allocation.
- **Enhanced Risk Management:** Predictive models for loan approvals and statuses enable banks to assess the risk profiles of applicants more accurately, reducing the chances of default and minimizing financial losses.
- **Fraud Prevention**: Real-time fraud detection systems can significantly reduce the incidence of fraudulent activities, protecting both the bank's assets and customers' funds.
- **Customer Retention**: Predictive analytics can identify at-risk customers, allowing banks to tailor retention strategies and improve customer loyalty.
- **Operational Efficiency:** Automating these processes reduces the need for manual intervention, speeding up decision-making and lowering operational costs**.**
- **Competitive Advantage**: Banks that adopt advanced machine learning techniques can gain a competitive edge by offering better services and more secure financial products.

2. Problem Statement:

The financial sector, particularly banking, has witnessed an exponential growth in data generation due to digitization. With vast amounts of data available, banks face significant challenges in analyzing and leveraging this data to improve decision-making processes and enhance customer experiences. Specifically, there are four critical areas where machine learning can be pivotal:

- **Loan Approval Prediction:** Determining whether a loan application should be approved based on the applicant's financial history and current status. These models can identify patterns and correlations that may not be immediately apparent to human analysts, enabling more accurate and fair loan approval decisions
- **Loan Status Prediction:** Forecasting the likelihood of a borrower defaulting on a loan or successfully repaying it. Machine learning algorithms can process historical loan data, repayment behavior, economic indicators, and borrower characteristics to forecast loan outcomes.
- **Credit Card Fraud Detection:** Identifying fraudulent transactions in realtime to prevent financial losses. Machine learning models can analyze transaction patterns, geographical locations, spending behaviors, and other contextual information to identify anomalies indicative of fraud.
- **Bank Customer Churn Prediction:** Predicting which customers are likely to close their accounts, allowing banks to take proactive measures to retain them.
- **Model Deployment:** Deploy the best-performing model to make predictions on new data. The deployed model should be scalable and capable of handling real-time predictions.

2.1 Data Set Information:

- Our banking project likely utilizes a dataset containing customer information relevant for various tasks like fraud detection, loan approval prediction, or customer churn prediction. This dataset would be structured in a similar tabular format, with each row representing a customer and each column representing a specific feature about the customer or their banking activity. The dataset contains the following 13 columns (features), which are used to train the machine learning models:

- Loan_ID
- Gender
- Married
- Dependents

- Education
- Self_Employed
- ApplicantIncome
- CoapplicantIncome
- LoanAmount
- Loan_Amount_Term
- Credit_History
- Property_Area
- Loan_Status

This comprehensive dataset, enriched with diverse features about our customers and their banking behavior, empowers machine learning models to accomplish your project's objectives. These objectives could be predicting fraud, making informed loan approval decisions, or retaining valuable customers by anticipating their needs and preferences.

3. Proposed Design:

Here is proposed design section for Banking and Insurance prediction:

- **Data Source:** The dataset is sourced from publicly available repositories Kaggle. Load the loan dataset from a CSV file using Pandas library.
- **Data Cleaning:** Handle any missing or duplicate values (if any). Convert categorical features to numeric values using techniques like one-hot encoding or label encoding.
- **Feature Scaling:** Normalize or standardize features to ensure they are on a similar scale. This is particularly important for algorithms sensitive to feature scales (e.g., logistic regression, SVM).
- **Train-Test Split:** Split the dataset into training and testing sets, typically using 75- 25 split, to evaluate the model's performance on unseen data.
- **Data Visualization:** Use histograms visualize feature distributions and identify potential outliers. Generate a correlation matrix and visualize it using a heatmap to understand the relationships between features and the target variable.
- **Model Selection:** Choose multiple classification models to compare their performance.

  Common models include:

  1.Logistic Regression

  2.Support Vector Machine

3. Decision Tree

- **Model Training:** Train each model on the training dataset. Use cross-validation to tune hyperparameters and avoid overfitting.
- **Evaluation Metrics:** Evaluate models using metrics such as accuracy, precision, recall, F1-score and support.
- **Confusion Matrix:** Generate confusion matrices to visualize the performance of each model in terms of true positives, false positives, true negatives, and false negatives.
- **Model Selection:** Select the best-performing model based on the evaluation metrics.
- **Project Report:** Document the entire process, including data collection, preprocessing, model training, evaluation, and deployment.

3.1 Libraries Used:

A. NumPy:

a. Provides support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

b. Offers efficient numerical operations, enabling faster data analysis.

B. Pandas:

a. Provides high-performance, easy-to-use data structures and data analysis tools for working with structured (tabular, multidimensional, potentially heterogeneous) and time series data.

b. Enables efficient data cleaning, preprocessing, and manipulation operations.

C. Matplotlib:

a. A comprehensive library for creating static, animated, and interactive visualizations in Python

b. Produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms.

D. Seaborn:

a. A data visualization library based on matplotlib, providing a high-level interface for drawing attractive and informative statistical graphics.

b. Offers a wide range of visualizations, including scatter plots, line plots, bar plots, and more.

E. Scikit-learn:

- **a.** A machine learning library that features various classification, regression, and clustering algorithms, as well as tools for model evaluation and selection.
- **b.** Provides efficient implementations of popular machine learning algorithms, such as Random Forest, Support Vector Machines, and Logistic Regression.

3.2. Methods Used:

1. **pd.read_csv():**

Reads a comma-separated values (CSV) file into a Pandas Data Frame.

2. **df.info():**

This method in pandas provides a concise summary of a Data Frame i.e. type of data frame

Object, index range, column information.

3. **df.isnull().sum():** This function is used to identify missing value in data frame.
4. **df.corr():** This function in pandas compute pairwise correlation of columns, excluding NA/null values.
5. **df.hist():** It is used to create histograms of data frame's numerical columns. It represents the distribution of data across different intervals or "bins".
6. **df.drop():** This method in pandas is used to remove rows or columns from a data frame.
7. **train_test_split():** This method is used to split a dataset into training and testing sets.
8. **StandardScaler():** It is a tool for standardizing features in a dataset before fitting a machine learning model. Standardization involves scaling the features to have a mean of 0 and standard deviation of 1.
9. **confusion_matrix():** It summarizes the number of correct and incorrect predictions made by the model on a set of test data.

10. **sns.heatmap():** It is used for creating heatmaps. Heatmaps are particularly useful for displaying matrices where the values encode some magnitude or relationship.

11. **accuracy score():** Computes the accuracy score of the model's predictions against true labels.

4. Results:

1. Imported required libraries such as pandas, numpy, seaborn, matplotlib, scikit-learn, and pickle.

2. Read the CSV file 'loan.csv' into a DataFrame called df.

3. Explored the data using functions like head(), tail(), shape, columns, info(), and isnull().sum().

4. Handled missing values by filling them with appropriate methods for numerical and categorical columns.

5. Conducted data visualization using seaborn to understand the distribution and relationships between variables, such as count plots and correlation heatmap.

6. Engineered new features like 'Total_Income' and applied log transformations to numerical features.

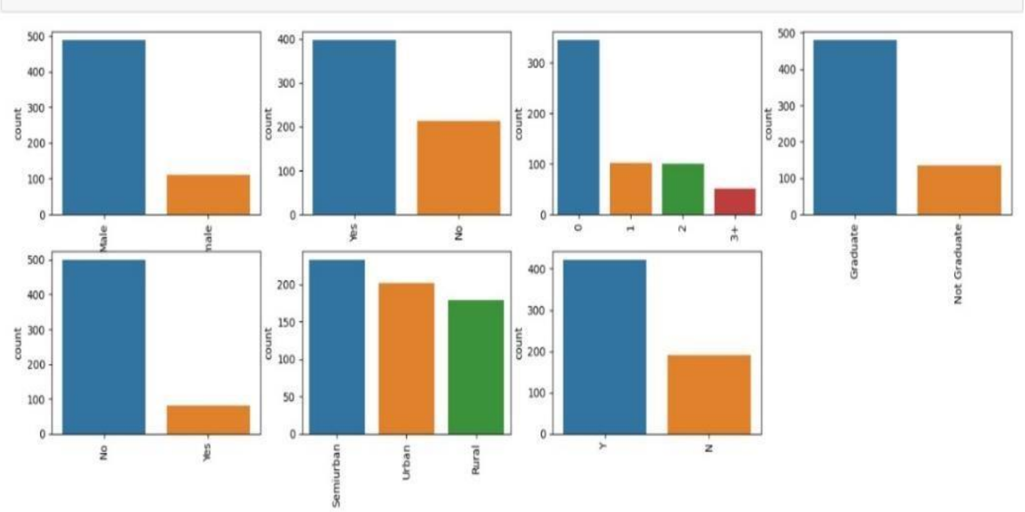
7. Encoded categorical variables using Label Encoding for binary columns and One-Hot Encoding for multi-class columns.

8. Split the data into independent features (X) and the target variable (y).

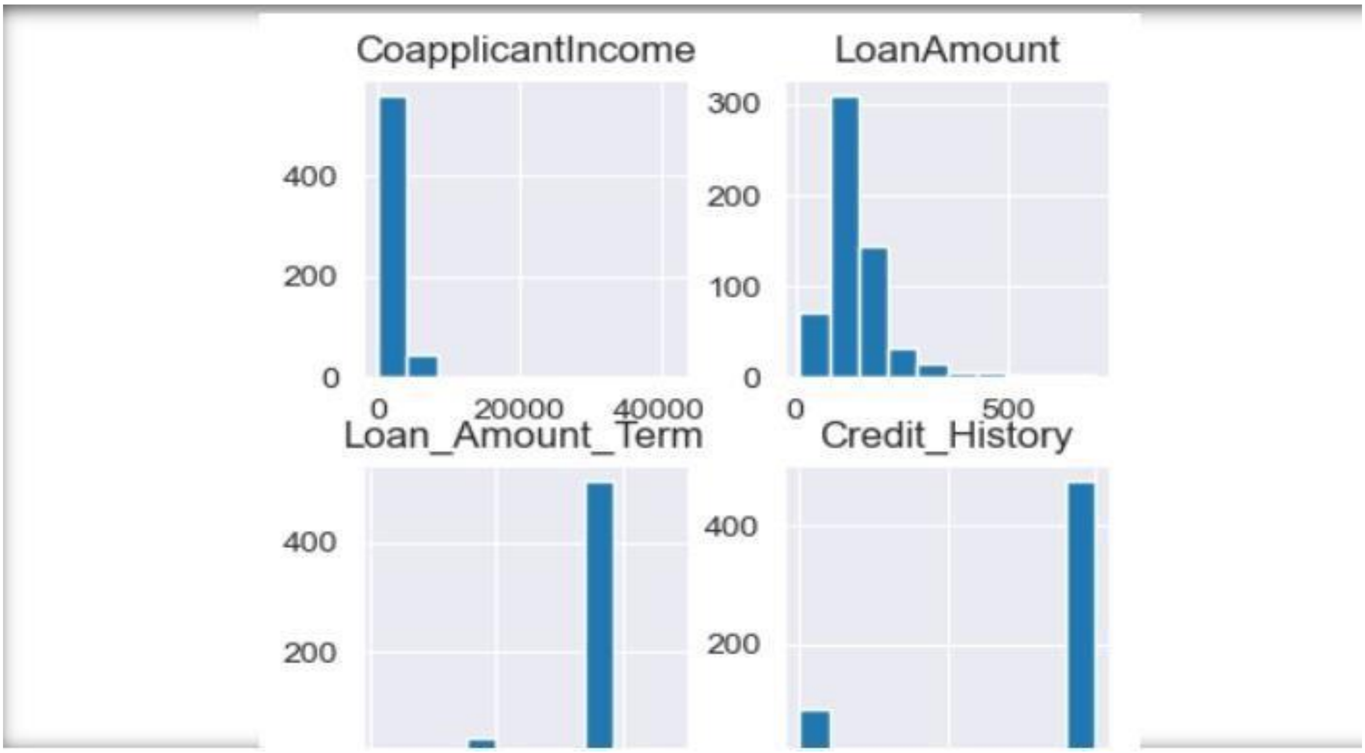
Conclusion: Following comprehensive evaluation and analysis, the Decision Tree model emerged victorious, surpassing Logistic Regression and Support Vector Machine models in accuracy and predictive power. Users can confidently rely on the Decision Tree for highly accurate predictions of banking prediction presence or absence.

5. Project Screenshots:

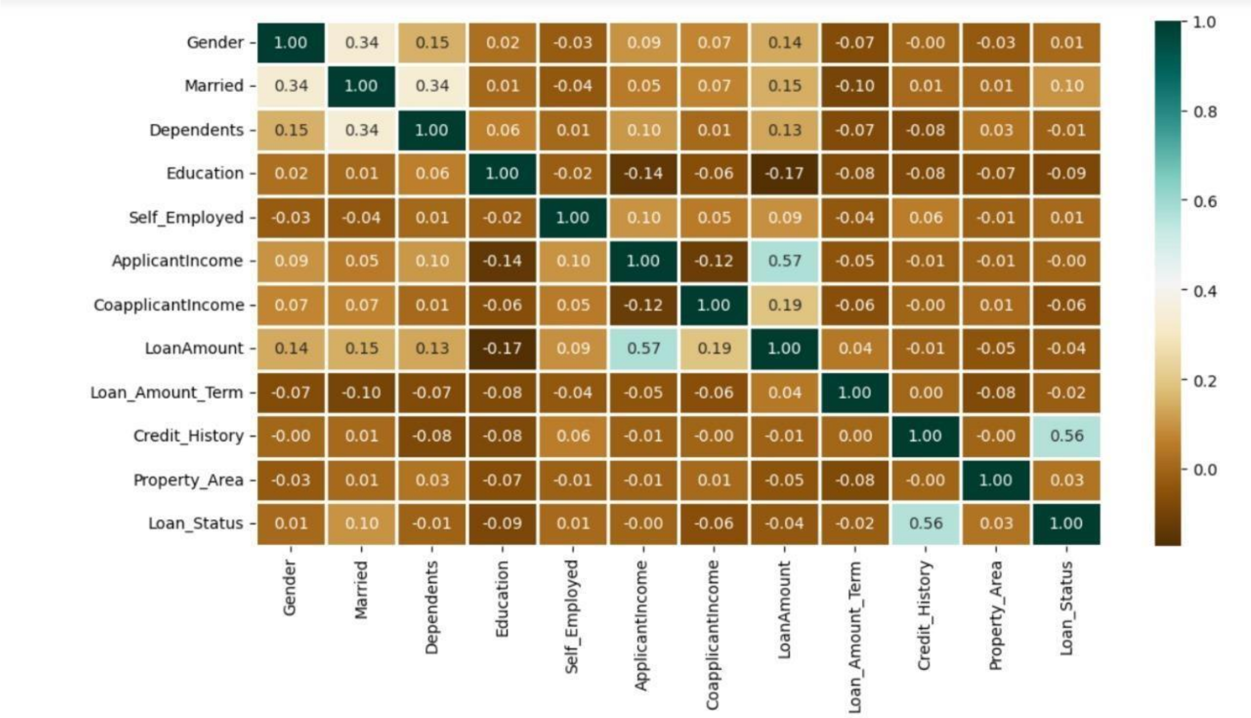
- Data Processing and Visualising:



- Incomes based on the features:

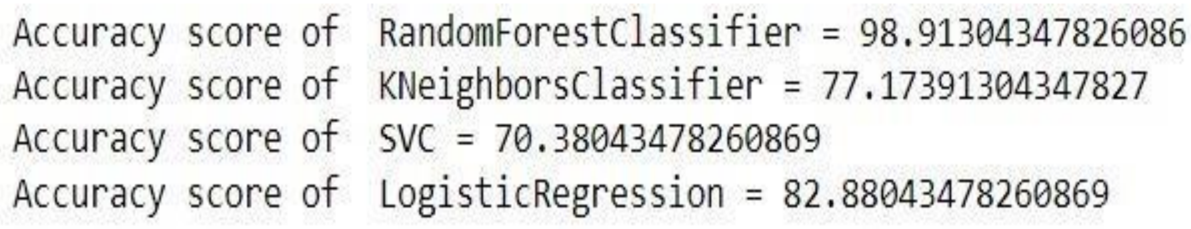


- HeatMap based on Features:



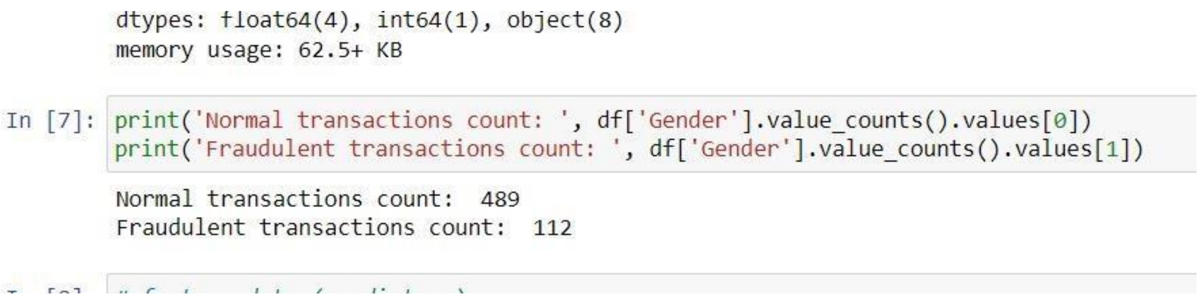
	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
Gender	1.00	0.34	0.15	0.02	-0.03	0.09	0.07	0.14	-0.07	-0.00	-0.03	0.01
Married	0.34	1.00	0.34	0.01	-0.04	0.05	0.07	0.15	-0.10	0.01	0.01	0.10
Dependents	0.15	0.34	1.00	0.06	0.01	0.10	0.01	0.13	-0.07	-0.08	0.03	-0.01
Education	0.02	0.01	0.06	1.00	-0.02	-0.14	-0.06	-0.17	-0.08	-0.08	-0.07	-0.09
Self_Employed	-0.03	-0.04	0.01	-0.02	1.00	0.10	0.05	0.09	-0.04	0.06	-0.01	0.01
ApplicantIncome	0.09	0.05	0.10	-0.14	0.10	1.00	-0.12	0.57	-0.05	-0.01	-0.01	-0.00
CoapplicantIncome	0.07	0.07	0.01	-0.06	0.05	-0.12	1.00	0.19	-0.06	-0.00	0.01	-0.06
LoanAmount	0.14	0.15	0.13	-0.17	0.09	0.57	0.19	1.00	0.04	-0.01	-0.05	-0.04
Loan_Amount_Term	-0.07	-0.10	-0.07	-0.08	-0.04	-0.05	-0.06	0.04	1.00	0.00	-0.08	-0.02
Credit_History	-0.00	0.01	-0.08	-0.08	0.06	-0.01	-0.00	-0.01	0.00	1.00	-0.00	0.56
Property_Area	-0.03	0.01	0.03	-0.07	-0.01	-0.01	0.01	-0.05	-0.08	-0.00	1.00	0.03
Loan_Status	0.01	0.10	-0.01	-0.09	0.01	-0.00	-0.06	-0.04	-0.02	0.56	0.03	1.00

- Accuracy of Loan Prediction:



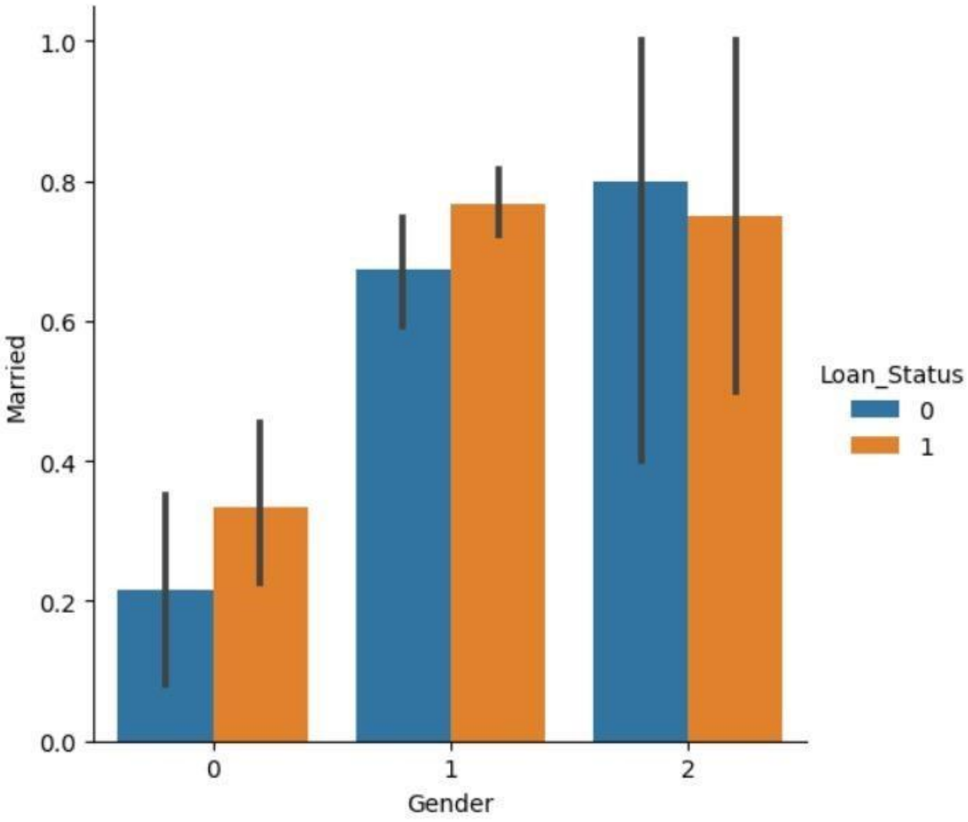
```
Accuracy score of  RandomForestClassifier = 98.91304347826086
Accuracy score of  KNeighborsClassifier = 77.17391304347827
Accuracy score of  SVC = 70.38043478260869
Accuracy score of  LogisticRegression = 82.88043478260869
```

- Fraud Detection:

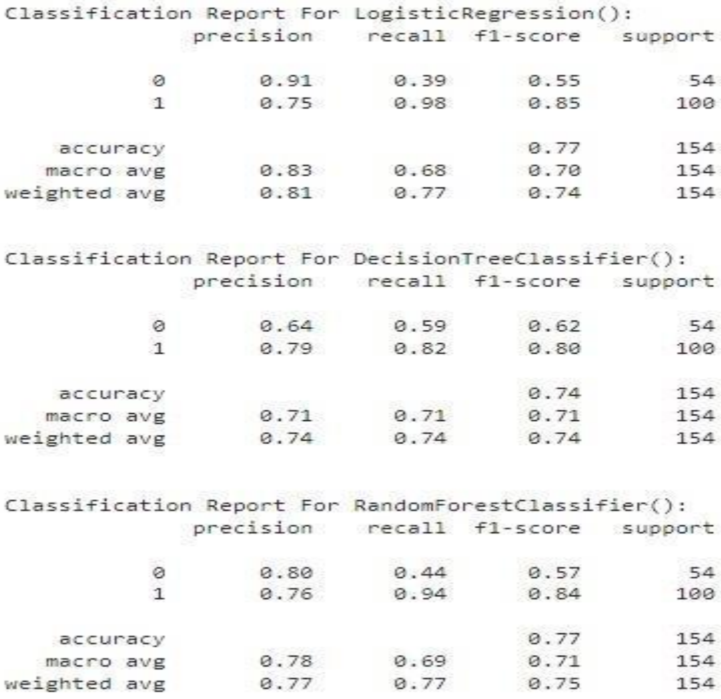


```
dtypes: float64(4), int64(1), object(8)
memory usage: 62.5+ KB

In [7]: print('Normal transactions count: ', df['Gender'].value_counts().values[0])
        print('Fraudulent transactions count: ', df['Gender'].value_counts().values[1])

Normal transactions count:  489
Fraudulent transactions count:  112
```

- Loan Status Prediction:

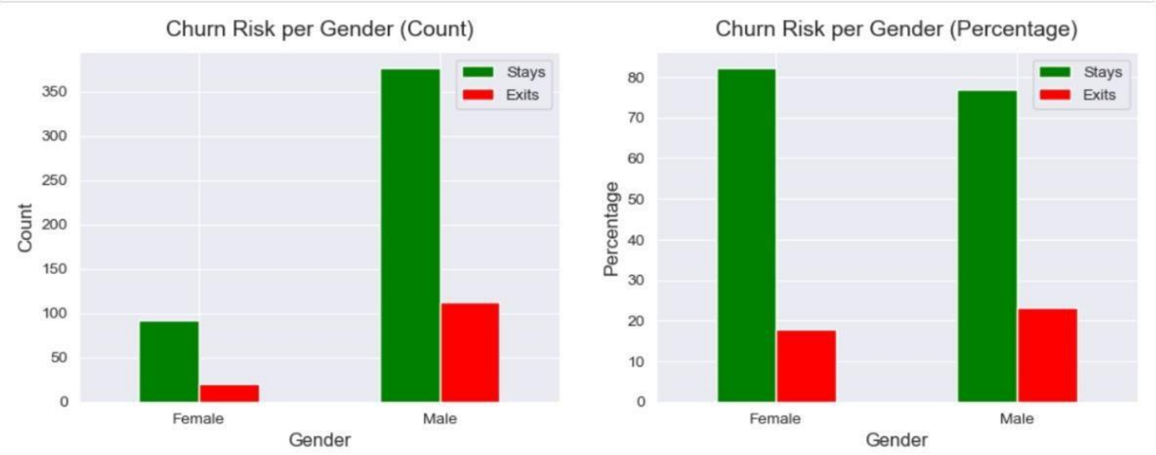


- Report of all classifiers:

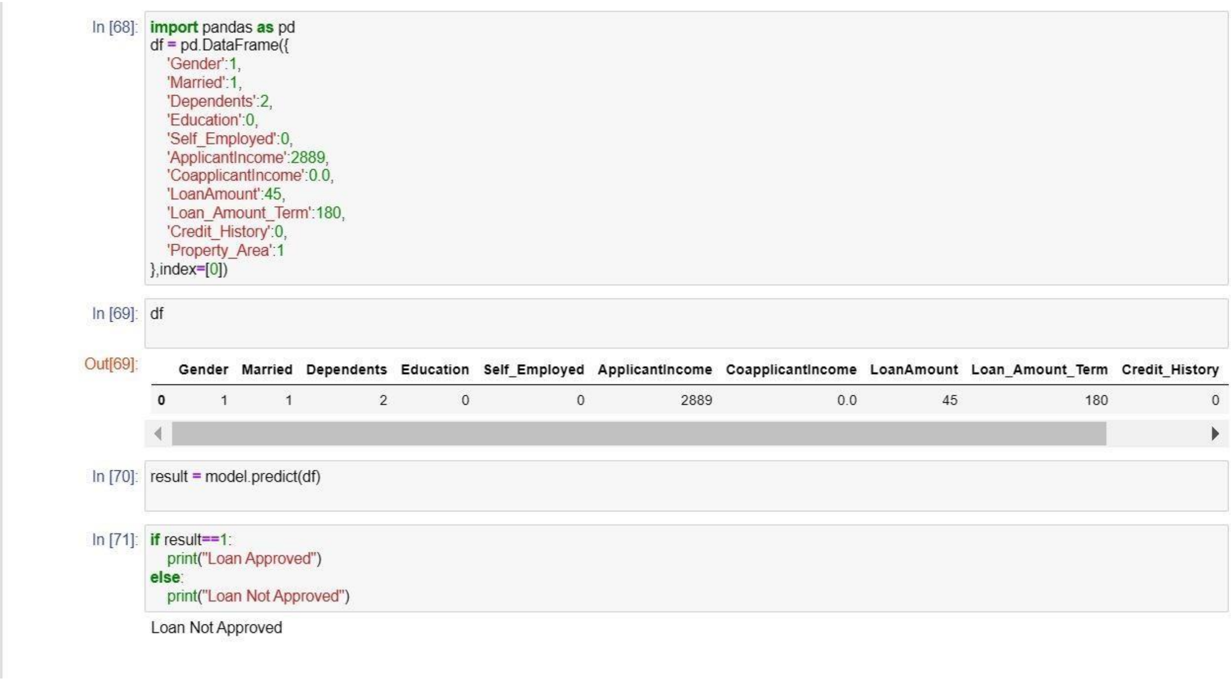


```
Classification Report For LogisticRegression():
              precision    recall  f1-score   support

           0       0.91      0.39      0.55        54
           1       0.75      0.98      0.85       100

    accuracy                           0.77       154
   macro avg       0.83      0.68      0.70       154
weighted avg       0.81      0.77      0.74       154

Classification Report For DecisionTreeClassifier():
              precision    recall  f1-score   support

           0       0.64      0.59      0.62        54
           1       0.79      0.82      0.80       100

    accuracy                           0.74       154
   macro avg       0.71      0.71      0.71       154
weighted avg       0.74      0.74      0.74       154

Classification Report For RandomForestClassifier():
              precision    recall  f1-score   support

           0       0.80      0.44      0.57        54
           1       0.76      0.94      0.84       100

    accuracy                           0.77       154
   macro avg       0.78      0.69      0.71       154
weighted avg       0.77      0.77      0.75       154
```

- Bank Customer Churn Prediction:



- Loan Approved or Not Approved:



```
In [68]: import pandas as pd
         df = pd.DataFrame({
             'Gender':1,
             'Married':1,
             'Dependents':2,
             'Education':0,
             'Self_Employed':0,
             'ApplicantIncome':2889,
             'CoapplicantIncome':0.0,
             'LoanAmount':45,
             'Loan_Amount_Term':180,
             'Credit_History':0,
             'Property_Area':1
         },index=[0])

In [69]: df

Out[69]:
     Gender  Married  Dependents  Education  Self_Employed  ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  Credit_History
  0       1        1           2          0              0             2889                0.0          45               180               0

In [70]: result = model.predict(df)

In [71]: if result==1:
             print("Loan Approved")
         else:
             print("Loan Not Approved")

         Loan Not Approved
```

6. Reference/Links:

1. Kaggle:
   https://www.kaggle.com/code/yonatanrabinovich/loan-prediction-dataset-ml-project
2. Geeks For Geeks: https://www.geeksforgeeks.org/supervised-machine-learning/
   https://www.geeksforgeeks.org/understanding-logistic-regression/
   https://www.geeksforgeeks.org/support-vector-machine-algorithm/
   https://www.geeksforgeeks.org/decision-tree/