Bank Loan Status prediction using ML Techniques

Team members:
Jayalakshmi Gunji
Priyanka Akula
Somna Sattoor
Naveen Yadav Gongati

Group-2

Project objective & motivation

- Optimize loan approval workflows to expedite processing times and enhance transparency throughout financial transactions.

- Proactively detect high-risk loans using predictive analytics to improve decision-making accuracy and reduce the incidence of loan defaults.

- Integrate real-time data processing to keep loan assessments up-to-date and reflective of current financial contexts, improving financial service responsiveness.

- Strengthen bank decision-making through advanced predictive modeling to minimize financial risks.

DATASET

Loan Status Classification
Data Source: https://www.kaggle.com/datasets/zaurbegiev/my-dataset/data
Data Shape: 100514 x 19
Target : Loan Status
Sample Dataset:

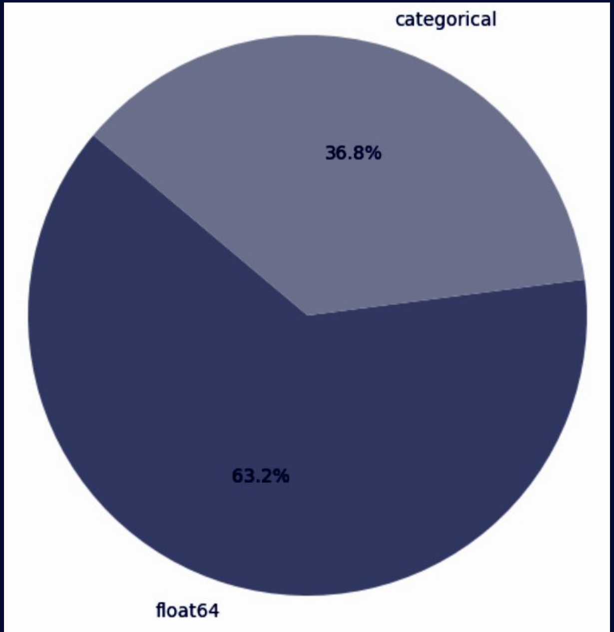
	Loan ID	Customer ID	Loan Status	Current Loan Amount	Term	Credit Score	Annual Income	Years in current job	Home Ownership	Purpose	Monthly Debt	Years of Credit History	Months since last delinquent	Number of Open Accounts	N of Pro
0	14dd8831-6af5-400b-83ec-68e61888a048	981165ec-3274-42f5-a3b4-d104041a9ca9	Fully Paid	445412.0	Short Term	709.0	1167493.0	8 years	Home Mortgage	Home Improvements	5214.74	17.2	NaN	6.0	
1	4771cc26-131a-45db-b5aa-537ea4ba5342	2de017a3-2e01-49cb-a581-08169e83be29	Fully Paid	262328.0	Short Term	NaN	NaN	10+ years	Home Mortgage	Debt Consolidation	33295.98	21.1	8.0	35.0	
2	4eed4e6a-aa2f-4c91-8651-ce984ee8fb26	5efb2b2b-bf11-4dfd-a572-3761a2694725	Fully Paid	99999999.0	Short Term	741.0	2231892.0	8 years	Own Home	Debt Consolidation	29200.53	14.9	29.0	18.0	
3	77598f7b-32e7-4e3b-a6e5-06ba0d98fe8a	e777faab-98ae-45af-9a86-7ce5b33b1011	Fully Paid	347666.0	Long Term	721.0	806949.0	3 years	Own Home	Debt Consolidation	8741.90	12.0	NaN	9.0	
4	d4062e70-befa-4995-8643-a0de73938182	81536ad9-5ccf-4eb8-befb-47a4d608658e	Fully Paid	176220.0	Short Term	NaN	NaN	5 years	Rent	Debt Consolidation	20639.70	6.1	NaN	15.0	

LITERATURE REVIEW

PAPER TITLE	AUTHOR	SUMMARY	YEAR
Prediction of Loan Status in Commercial Bank using Machine Learning Classifier	G. Arutjothi, Dr. C. Senthamarai	This paper proposes a machine learning model using K-Nearest Neighbor (K-NN) classifier combined with Min-Max normalization to predict loan status in commercial banks, aiming to improve accuracy in classifying credit defaulters.	2017
Machine Learning Models for Predicting Bank Loan Eligibility	Ugochukwu .E. Orji, Chikodili .H. Ugwuishiwu, Joseph. C. N. Nguemaleu, Peace. N. Ugwuanyi	This paper explores using six ML algorithms (Random Forest, Gradient Boost, Decision Tree, SVM, KNN, and Logistic Regression) for predicting loan eligibility, highlighting that Random Forest had the highest performance accuracy.	2022
Logistic Regression Model for Loan Prediction: A Machine Learning Approach	Richa Manglani, Anuja Bokhare	This study uses logistic regression to predict loan approval. The method simplifies the loan approval process, making it quicker and more efficient by automating the evaluation of applicant features.	
A Combination Method of Resampling and Random Forest for Imbalanced Data Classification	Liu Zheng, Qiu Han, Zhu Junhu	This paper proposes a combination of resampling and Random Forest techniques to handle imbalanced datasets in applications like credit card fraud detection, enhancing the classification performance of minority classes.	2022
Loan Eligibility Prediction using Machine Learning based on Personal Information	M. Meenaakumari, Dr. Seema Sharma, P. Jayasuriya, Geetha Manoharan, Dr. Nasa Dhanraj, Mohit Tiwari	This paper develops a machine learning model to predict health loan eligibility using Random Forest, Naive Bayes, and Linear Regression algorithms, finding Random Forest to perform the best in terms of accuracy and error.	2022

EDA

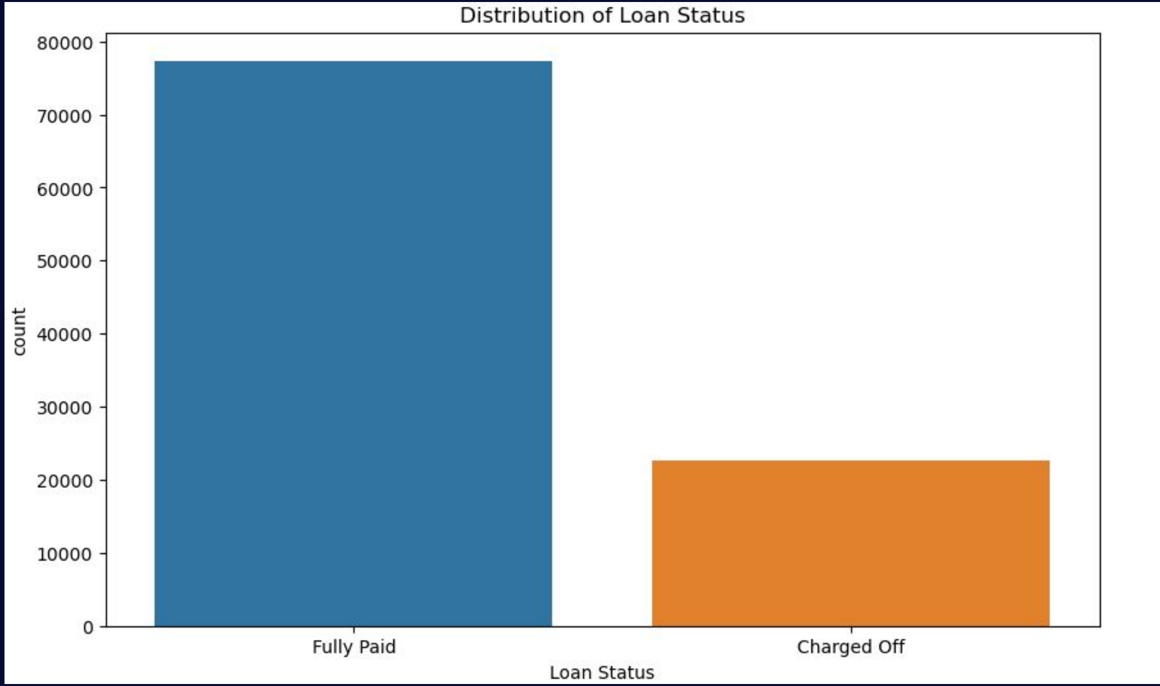
Data Type Distribution




```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100514 entries, 0 to 100513
Data columns (total 19 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   Loan ID                       100000 non-null  object
 1   Customer ID                   100000 non-null  object
 2   Loan Status                   100000 non-null  object
 3   Current Loan Amount           100000 non-null  float64
 4   Term                          100000 non-null  object
 5   Credit Score                  80846 non-null   float64
 6   Annual Income                 80846 non-null   float64
 7   Years in current job          95778 non-null   object
 8   Home Ownership                100000 non-null  object
 9   Purpose                       100000 non-null  object
 10  Monthly Debt                  100000 non-null  float64
 11  Years of Credit History       100000 non-null  float64
 12  Months since last delinquent  46859 non-null   float64
 13  Number of Open Accounts       100000 non-null  float64
 14  Number of Credit Problems     100000 non-null  float64
 15  Current Credit Balance        100000 non-null  float64
 16  Maximum Open Credit           99998 non-null   float64
 17  Bankruptcies                  99796 non-null   float64
 18  Tax Liens                     99990 non-null   float64
dtypes: float64(12), object(7)
memory usage: 14.6+ MB
```

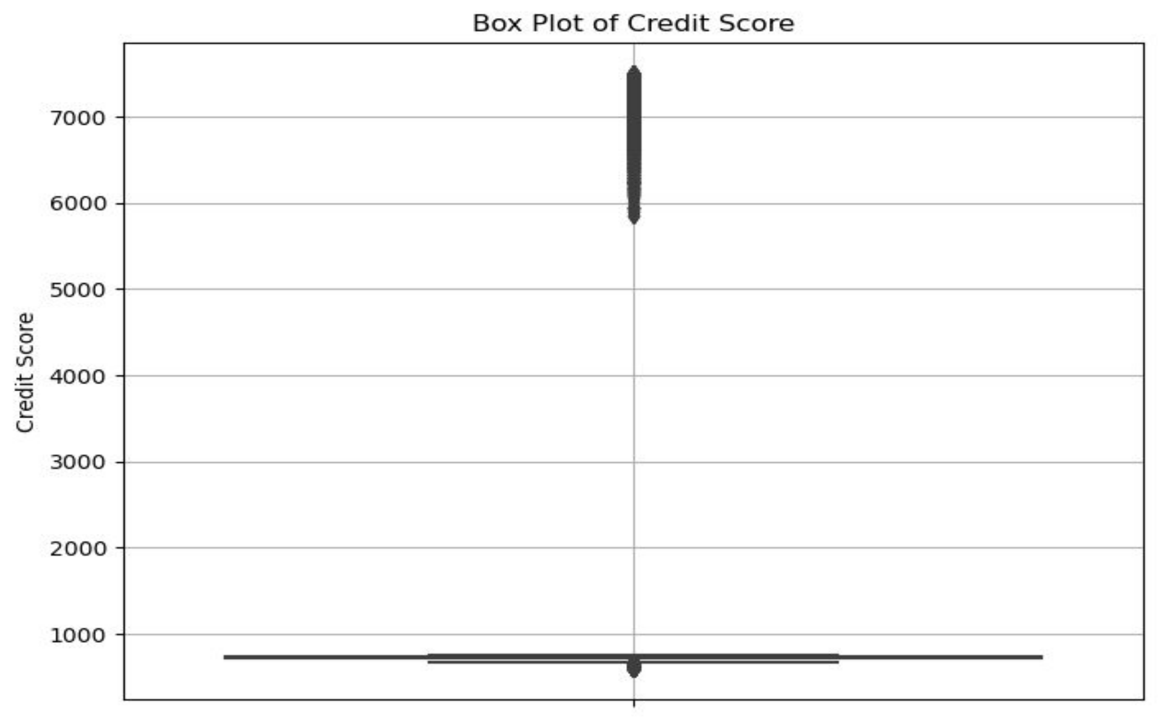
EDA

Distribution of Loan Status



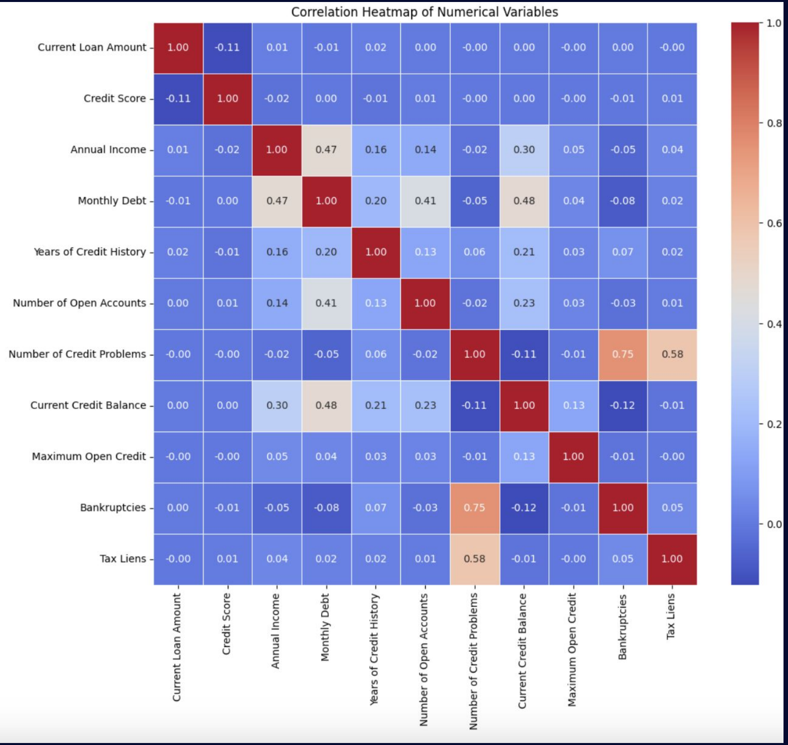
EDA

Boxplot of Credit Score



EDA

Correlation Matrix between features



DATA PREPROCESSING

Data Cleaning:
Removed rows with missing features
Removed features with more than 80% missing values.

Outliers:
Removed the outliers based on the credit score range (300,850) using IQR.

Correlation Analysis:
Checked for correlation between Features and dropped columns Tax Liens and Monthly debt.

Steps	Datasize
Raw Dataset	(100514 x 19)
3 Columns dropped	(100514 x 16)
Duplicates found and dropped.	(89789 x 16)
Outliers found and dropped	(76295 x 16)
missing values found and dropped.	(67493 x 16)
Shape after Data Cleaning	(67490 x 16)

DATA PREPROCESSING

One hot encoding:
Except target variable converted categorical variables into a numerical format

```
Categorical columns before excluding target: Index(['Loan Status', 'Term', 'Years in current job', 'Home Ownership',
       'Purpose'],
      dtype='object')
Categorical columns after excluding target: Index(['Term', 'Years in current job', 'Home Ownership', 'Purpose'], dtype='object')
Applied One-Hot Encoding. Columns now: Index(['Loan Status', 'Current Loan Amount', 'Credit Score', 'Annual Income',
       'Monthly Debt', 'Years of Credit History', 'Number of Open Accounts',
       'Number of Credit Problems', 'Current Credit Balance',
       'Maximum Open Credit', 'Bankruptcies', 'Tax Liens', 'Term_Long Term',
       'Term_Short Term', 'Years in current job_1 year',
       'Years in current job_10+ years', 'Years in current job_2 years',
       'Years in current job_3 years', 'Years in current job_4 years',
       'Years in current job_5 years', 'Years in current job_6 years',
       'Years in current job_7 years', 'Years in current job_8 years',
       'Years in current job_9 years', 'Years in current job_< 1 year',
       'Home Ownership_HaveMortgage', 'Home Ownership_Home Mortgage',
       'Home Ownership_Own Home', 'Home Ownership_Rent',
       'Purpose_Business Loan', 'Purpose_Buy House', 'Purpose_Buy a Car',
       'Purpose_Debt Consolidation', 'Purpose_Educational Expenses',
       'Purpose_Home Improvements', 'Purpose_Medical Bills', 'Purpose_Other',
       'Purpose_Take a Trip', 'Purpose_major_purchase', 'Purpose_moving',
       'Purpose_other', 'Purpose_renewable_energy', 'Purpose_small_business',
       'Purpose_vacation', 'Purpose_wedding'],
      dtype='object')
Data after encoding:        Loan Status  Current Loan Amount  Credit Score  Annual Income  ...  Purpose_renewable_energy  Purpose_small_business  Purpose_vacation  Purpos
e_wedding
0       Fully Paid             445412.0         709.0      1167493.0  ...                    False                   False              False            False
2       Fully Paid           99999999.0         741.0      2231892.0  ...                    False                   False              False            False
3       Fully Paid             347666.0         721.0       806949.0  ...                    False                   False              False            False
5      Charged Off             206602.0        7290.0       896857.0  ...                    False                   False              False            False
6       Fully Paid             217646.0         730.0      1184194.0  ...                    False                   False              False            False
...            ...                  ...           ...            ...  ...                      ...                     ...                ...              ...
99990   Fully Paid           99999999.0         742.0      1190046.0  ...                    False                   False              False            False
99994   Fully Paid             210584.0         719.0       783389.0  ...                    False                   False              False            False
99996   Fully Paid           99999999.0         732.0      1289416.0  ...                    False                   False              False            False
99997   Fully Paid             103136.0         742.0      1150545.0  ...                    False                   False              False            False
99998   Fully Paid             530332.0         746.0      1717524.0  ...                    False                   False              False            False

[67490 rows x 45 columns]
```

DATA PREPROCESSING

SMOTE analysis on Target variable

SMOTE analysis:
Generated synthetic samples for the Target Variable

```
Data split into training and testing sets. Training shape: (48485, 44), Testing shape: (12122, 44)
Class distribution before SMOTE:
Loan Status
1    39767
0     8718
Name: count, dtype: int64
Applied SMOTE. Balanced training data shape: (79534, 44)
Class distribution after SMOTE:
Loan Status
1    39767
0    39767
Name: count, dtype: int64
Data before scaling:
   Current Loan Amount  Credit Score  Annual Income  Monthly Debt  ...  Purpose_renewable_energy  Purpose_small_business  Purpose_vacation  Purpose_wedding
0             269126.0         736.0       871587.0      10822.21  ...                     False                   False             False            False
1             331562.0         743.0      1336251.0      11469.35  ...                     False                   False             False            False
2           99999999.0         737.0       652897.0      15343.07  ...                     False                   False             False            False
3             177870.0         731.0       825664.0       8187.86  ...                     False                   False             False            False
4             142230.0         726.0       671726.0       5412.91  ...                     False                   False             False            False
```

DATA PREPROCESSING

Feature Scaling:
StandardScaler standardizes features by removing the mean and scaling to unit variance.

Feature Scaling

```
Data before scaling:
   Current Loan Amount  Term  ...  Bankruptcies  Tax Liens
0              43142.0     1  ...           0.0        0.0
1             545600.0     1  ...           0.0        0.0
2             394548.0     1  ...           0.0        0.0
3             232430.0     1  ...           0.0        0.0
4           99999999.0     1  ...           0.0        0.0

[5 rows x 15 columns]
Data after scaling:
[[-0.35663791  0.76897403 -0.4262926  -0.51593105 -0.87399208 -0.95526629
  -0.34366072  0.0304761   0.14030968  0.3590117  -0.36112044 -0.16493497
  -0.01252135 -0.34297799 -0.12408651]
 [-0.34038665  0.76897403 -0.42012448  0.12363472 -0.87399208  1.23141493
  -0.34366072  1.2059969   0.12483832 -1.29562791 -0.36112044  0.01082073
   0.03052127 -0.34297799 -0.12408651]
 [-0.3452722   0.76897403 -0.43451676  0.78294629 -0.87399208 -0.95526629
  -0.34366072  0.98361147  0.3414374   0.15218174 -0.36112044  0.16957611
   0.0123013  -0.34297799 -0.12408651]
 [-0.35051567  0.76897403 -0.42423656  0.95325386 -0.20503355 -0.95526629
  -0.34366072  1.50204579  0.23313786 -0.88196801 -0.36112044 -0.01306191
  -0.05562492 -0.34297799 -0.12408651]
 [ 2.87631749  0.76897403 -0.4314327  -0.06519509 -0.87399208 -0.95526629
  -0.34366072  0.91963297  1.0840628  -0.46830811 -0.36112044  2.07145621
   0.09104914 -0.34297799 -0.12408651]]
```

EXPERIMENTS AND RESULTS

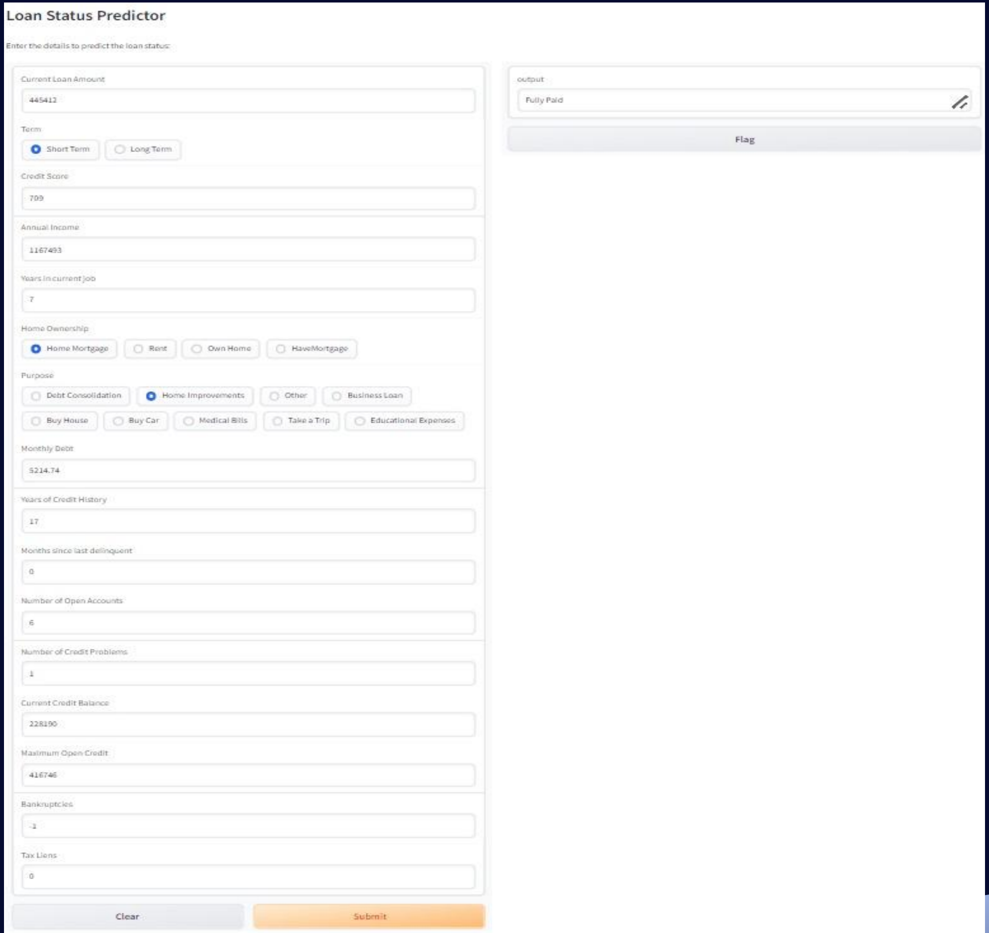
Model Type	Model	Baselines	Hyperparameter Tuning	Best Parameters
Plain ML Models	Random Forest	Accuracy: 081923, Precision: 0.7839, Recall: 0.8652, F1 Score: 0.8201	Accuracy: 0.8618, Precision: 0.8324, Recall: 0.9073, F1 Score: 0.8683	max_features:'sqrt', n_estimators: 200
	XGBoost	Accuracy: 0.8207, Precision: 0.7990, Recall: 0.8953, F1 Score: 0.8448	Accuracy: 0.8638, Precision: 0.8091, Recall: 0.9536, F1 Score: 0.8754	learning_rate: 0.1, max_depth: 7, n_estimators: 200
	KNN	Accuracy: 0.7462, Precision: 0.7019, Recall: 0.7612, F1 Score: 0.7435	Accuracy: 0.7616, Precision: 0.7461, Recall: 0.7961, F1 Score: 0.7703	n_neighbors : 15
	Logistic Regression	Accuracy: 0.7293, Precision: 0.6984, Recall: 0.7961, F1 Score: 0.7398	Accuracy: 0.7439, Precision: 0.7131, Recall: 0.8196, F1 Score: 0.7627	C : 1

EXPERIMENTS AND RESULTS

Model Type	Model	Baselines	Best Parameters
Ensemble Technique	Random Forest, XGBoost, KNN and Logistic Regression	Accuracy: 0.8053, Precision: 0.9130, Recall: 0.6702, F1 Score: 0.7730	Best of all parameters

Application

Created Web application for Prediction



CONCLUSION

- Random Forest and XGBoost models outperformed logistic regression and KNN.
- Hyperparameter tuning and ensemble methods enhanced model performance further.

FUTURE SCOPE

- Implementing mechanisms to continuously update and retrain models with new data to adapt to changing patterns and improve prediction accuracy.

- Exploring additional features or creating new features based on domain knowledge to improve model performance.

- Utilizing clustering techniques to segment customers based on their financial behavior and preferences, enabling personalized lending solutions.

CONTRIBUTIONS

PROBLEM STATEMENT	Jayalakshmi , Priyanka, Somna, Naveen
LITERATURE SURVEY	Jayalakshmi, Somna
DATA COLLECTION	Priyanka, Naveen
DATA PRE-PROCESSING	Priyanka, Naveen , Jayalakshmi , Somna
FEATURE ENGINEERING	Priyanka, Naveen
MODELING	Priyanka, Naveen , Jayalakshmi , Somna
EXPERIMENT MODELING	Jayalakshmi, Somna
REPORT and PRESENTATION SLIDES	Jayalakshmi , Priyanka, Somna, Naveen

REFERENCES

- R. Manglani and A. Bokhare, "Logistic Regression Model for Loan Prediction: A Machine Learning Approach," *2021 Emerging Trends in Industry 4.0 (ETI 4.0)*, Raigarh, India, 2021, pp. 1-6, doi: 10.1109/ETI4.051663.2021.9619201.
- U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu and P. N. Ugwuanyi, "Machine Learning Models for Predicting Bank Loan Eligibility," 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), Lagos, Nigeria, 2022, pp. 1-5, doi: 10.1109/NIGERCON54645.2022.9803172.
- L. Zheng, Q. Han and Z. Junhu, "A Combination Method of Resampling and Random Forest for Imbalanced Data Classification," *2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC)*, Suzhou, China, 2022, pp. 1-5, doi: 10.1109/CTISC54888.2022.9849803.
- A. Mueankoo, M. Eso and S. Musikasuwan, "Performance of a Loan Repayment Status Model Using Machine Learning," *2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Prachuap Khiri Khan, Thailand, 2022, pp. 1-4, doi: 10.1109/ECTI-CON54298.2022.9795570.
- M. Meenaakumari, P. Jayasuriya, N. Dhanraj, S. Sharma, G. Manoharan and M. Tiwari, "Loan Eligibility Prediction using Machine Learning based on Personal Information," *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, Uttar Pradesh, India, 2022, pp. 1383-1387, doi: 10.1109/IC3I56241.2022.10073318.
- G. Arutjothi and C. Senthamarai, "Prediction of loan status in commercial bank using machine learning classifier," *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, Palladam, India, 2017, pp. 416-419, doi: 10.1109/ISS1.2017.8389442.

THANK YOU