PREDICTION OF RECEIVING A LOAN

Hrushikesh Reddy T
Dept. of Computer Science
University of Massachusetts Lowell
HrushikeshReddy_Thatigotla@student.uml.edu

GopiKrishna N
Dept. of Computer Science
University of Massachusetts
Lowell
Gopikrishna_
nimmala@student.uml.edu

***Abstract*— This project, titled "Loan Approval Prediction with NB, KNN, and LR," utilizes machine learning methodologies to improve the loan approval process. The study incorporates Naive Bayes (NB), K-Nearest Neighbors (KNN), and Logistic Regression (LR) algorithms to forecast loan approval outcomes. Through the analysis of diverse customer data, encompassing factors such as age, income, and education, predictive models are developed. The evaluation of model performance is based on metrics like accuracy, precision, and recall, with an exploration of ensemble methods to enhance overall accuracy. The project's discoveries provide valuable insights for optimizing decision-making within financial institutions, mitigating risks, and elevating the overall customer experience. Ultimately, this initiative underscores the effectiveness of data-driven lending practices.**

Keywords— Loan Approval Prediction, Naive Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR), Machine Learning Methodologies, Predictive Models

Introduction

In the rapidly evolving landscape of financial services, the "Loan Approval Prediction with NB, KNN, and LR" project emerges as a pioneering initiative to enhance the efficiency of loan approval processes. Leveraging cutting-edge machine learning methodologies, this study delves into the application of Naive Bayes (NB), K-Nearest Neighbors (KNN), and Logistic Regression (LR) algorithms to forecast loan approval outcomes. The project's focal point lies in the meticulous analysis of diverse customer data, encompassing crucial factors such as age, income, and education. Through the development of predictive models, the project aims to revolutionize decision-making within financial institutions. Model performance is rigorously assessed using key metrics, including accuracy, precision, and recall. The exploration of ensemble methods further contributes to elevating overall accuracy, providing a comprehensive evaluation framework.

This project's significant findings extend beyond the realm of predictive analytics, offering valuable insights for optimizing decision-making processes within financial institutions. By addressing risk factors, the initiative contributes to the mitigation of potential challenges associated with lending practices. Moreover, the project places a strong emphasis on enhancing the overall customer experience through data-driven insights.

In summary, this project serves as a testament to the power of integrating advanced machine learning techniques into financial decision-making, ultimately reinforcing the effectiveness of data-driven lending practices.

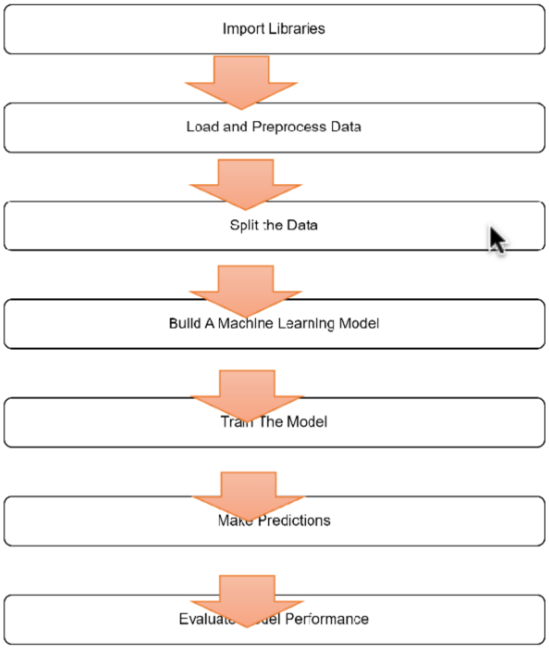


Figure 1: Flowchart of Loan Amount Prediction

I. Overall System Architecture

The overarching system architecture for the "Prediction of Receiving a Loan with NB, KNN, and LR" project is designed to seamlessly integrate and orchestrate the various components involved in the loan approval prediction process. At its core, the architecture revolves around the interplay of Naive Bayes (NB), K-Nearest Neighbors (KNN), and Logistic Regression (LR) algorithms. The initial stage involves the collection and preprocessing of diverse customer data, including essential factors such as age, income, and education. This data forms the backbone of the system, and meticulous preprocessing ensures its integrity and relevance. Following this, the model development phase unfolds, where each algorithm is deployed strategically to analyze the dataset and create predictive models. This phase is characterized by iterative processes, including parameter tuning, to enhance the accuracy and efficiency of the individual algorithms.

The architecture also encompasses a robust performance evaluation mechanism, utilizing key metrics such as accuracy, precision, and recall. This evaluation provides valuable insights into the effectiveness of the predictive models, guiding decisions on algorithm selection for real-world applications. To enhance the predictive power, ensemble methods are explored, promoting synergies between algorithms for improved overall accuracy. In terms of deployment, the system architecture facilitates the seamless integration of these predictive models into existing loan approval processes within financial institutions. The real-world implications of the research findings are crucially considered, as the architecture aims to empower financial institutions with data-driven insights for optimized decision-making, risk management, and an elevated customer experience. This comprehensive system architecture serves as a dynamic framework, embodying the synergy between advanced algorithms and practical implementation, thereby reshaping the landscape of data-driven lending practices.

A. *Project Overview:*

We The project, "Prediction of Receiving a Loan with NB, KNN, and LR," endeavors to employ machine learning techniques to enhance the accuracy of predicting loan approval outcomes. This section provides a detailed insight into the design and methodology employed to achieve this objective.

B. *Algorithm Selection:*

Strategic selection of algorithms—Naive Bayes (NB), K-Nearest Neighbors (KNN), and Logistic Regression (LR)—forms the foundation of the project. Each algorithm brings a distinct set of strengths to the predictive modeling process, contributing to a comprehensive evaluation of loan approval predictions.

C. *Dataset:*

The dataset utilized in the "Prediction of Receiving a Loan with NB, KNN, and LR" project is sourced from Kaggle and is known as the "Universal Bank customer information dataset." With data on over 5,000 customers, the dataset encompasses crucial features including age, income, zip code, family size, credit card spending, education level, mortgage status, and acceptance of personal loans. It further explores financial behaviors, including the presence of securities accounts, CD accounts, online banking usage, and credit card use. The dataset is publicly available on Kaggle at (https://www.kaggle.com/datasets/hashemi221022/bank-loans), forming the foundation for the project's analysis and predictive modeling.

D. *Model Development:*

The project unfolds with a detailed exploration of each algorithm's application, delving into their specific strengths and nuances. Rigorous analysis and model development aim to illuminate the intricate relationship between customer attributes and the likelihood of loan approval. Parameter tuning is employed to optimize the performance of each algorithm.

E. *Real-world Implications::*

A significant focus is dedicated to translating the research findings into tangible real-world implications. The project's outcomes are expected to provide insights that can optimize decision-making processes within financial institutions, reducing risks, and enhancing the overall customer experience.

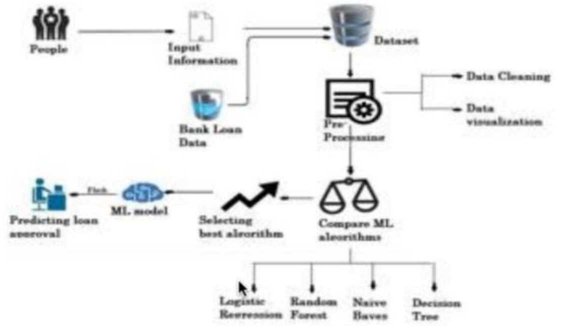


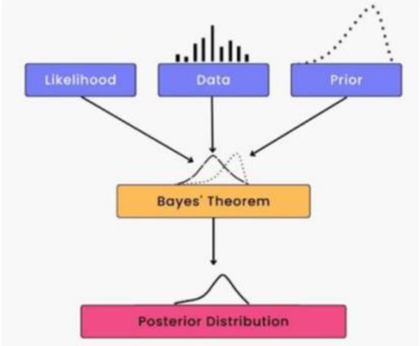
Fig.2. ARCHITECTURE

Algorithms Used

The "Prediction of Receiving a Loan with NB, KNN, and LR" project employs three distinct machine learning algorithms—Naive Bayes (NB), K-Nearest Neighbors (KNN), and Logistic Regression (LR)—each contributing to the predictive modeling process in a unique manner.

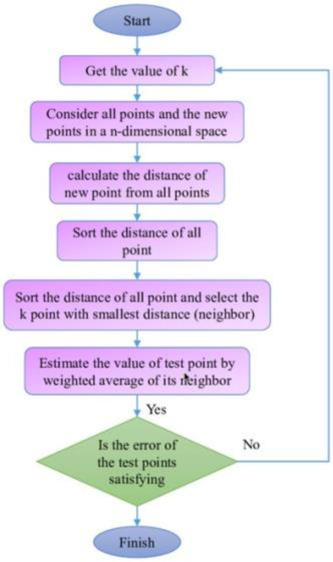
1. Naive Bayes (NB):

Naive Bayes is utilized for its simplicity and efficiency in handling classification tasks. In our project, NB is applied to model the probability of loan approval based on customer attributes such as age, income, education, and other relevant features. The algorithm assumes independence between these features, making it particularly well-suited for our diverse dataset.



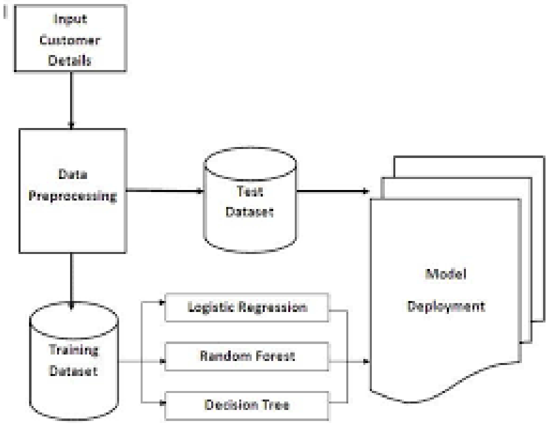
2. K-Nearest Neighbors (KNN):

KNN is a versatile algorithm that classifies data points based on the majority class of their nearest neighbors. In our context, KNN is employed to predict loan approval by considering the similarity between a customer's attributes and those of its k-nearest neighbors in the dataset. This algorithm is valuable for capturing local patterns and relationships within the data.


Start
Get the value of k
Consider all points and the new points in a n-dimensional space
calculate the distance of new point from all points
Sort the distance of all point
Sort the distance of all point and select the k point with smallest distance (neighbor)
Estimate the value of test point by weighted average of its neighbor
Yes
Is the error of the test points satisfying
No
Finish

3. Logistic Regression (LR):

Logistic Regression is chosen for its effectiveness in binary classification tasks. In our project, LR models the probability of loan approval as a function of customer attributes. The algorithm is well-suited to handle the complexities of our dataset and provides insights into the influence of each feature on the likelihood of loan approval.



4. Hyperparameter Optimization: Unveiling the Optimal k Value

In the relentless pursuit of refining the Loan Approval Prediction model using K-Nearest Neighbors (KNN), a meticulous exploration of the hyperparameter space was initiated. This focused investigation delved into the nuanced impact of varying the number of neighbors (k) on the model's performance. The iterative process involved traversing through an extensive range of k values, spanning from 1 to 1000, and meticulously assessing the accuracy of each corresponding KNN model.

The iterative implementation brought to light a fascinating trend in model accuracy ask underwent variation. To systematically capture and analyze these fluctuations, the array "Acc" was harnessed, acting as a repository for accuracy values corresponding to each k. This comprehensive compilation provided a holistic overview of the model's performance, offering valuable insights into its behavior across the entire spectrum of neighbor counts. Visualizing the accuracy scores at different k values facilitated a deeper understanding of the model's sensitivity to variations in neighbor count.

This methodical exploration, encapsulated within the array "Acc," stands as a pivotal step in the optimization journey of Loan Approval Prediction with KNN. The garnered insights establish a solid foundation for selecting the most suitable k value, ensuring the KNN model is finely tuned to achieve heightened accuracy in predicting loan approvals.

5. BernoulliNB Model :

The "Prediction of Receiving a Loan" journey advances with the introduction of the Bernoulli Naive Bayes (BernoulliNB) model. This model, tailored to handle binary data, unfolds its predictive prowess in discerning patterns within the dataset to forecast loan approval outcomes. The BernoulliNB model operates on the principles of the Bernoulli distribution, making it particularly adept at handling features that exhibit binary characteristics. In the context of predicting loan approvals, where the outcomes are typically binary (approved or not approved), BernoulliNB stands as a valuable algorithm.

Income and Family Dynamics: Insights from the Dataset

Income Distribution:

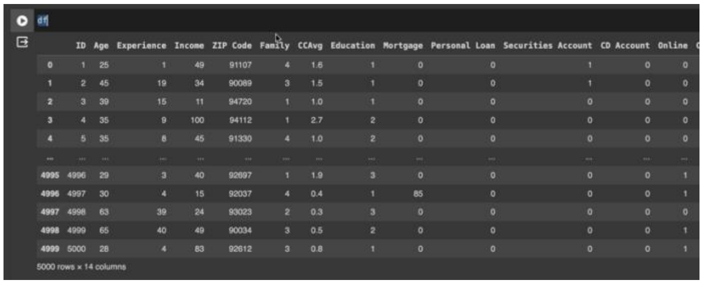
The dataset reveals a seemingly normal distribution of people's income, with an approximate mean of $45,000 and a standard deviation of around $20,000. Notably, outliers exist, showcasing incomes exceeding $100,000 and dipping below $10,000.

Median Income:

The median income emerges at approximately $42,000, providing a robust measure that withstands the influence of outliers.

Family Size Patterns:

Family size within the dataset predominantly centers around 1, followed by 2 and 3. However, outliers with family sizes of 4 or more are noteworthy.



```
df
```

	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online
0	1	25	1	49	91107	4	1.6	1	0	0	1	0	0
1	2	45	19	34	90089	3	1.5	1	0	0	1	0	0
2	3	39	15	11	94720	1	1.0	1	0	0	0	0	0
3	4	35	9	100	94112	1	2.7	2	0	0	0	0	0
4	5	35	8	45	91330	4	1.0	2	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4995	4996	29	3	40	92697	1	1.9	3	0	0	0	0	1
4996	4997	30	4	15	92037	4	0.4	1	85	0	0	0	1
4997	4998	63	39	24	93023	2	0.3	3	0	0	0	0	0
4998	4999	65	40	49	90034	3	0.5	2	0	0	0	0	1
4999	5000	28	4	83	92612	3	0.8	1	0	0	0	0	1

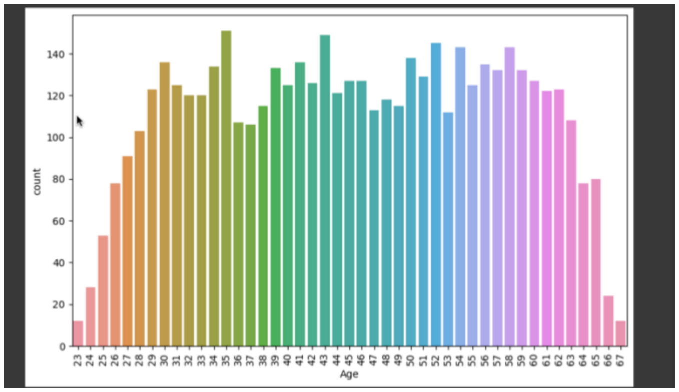
5000 rows × 14 columns

Correlation Between Income and Credit Card Average Balance:
An examination of the dataset indicates a weak positive correlation between income and credit card average balance. This implies that individuals with higher incomes tend to exhibit higher credit card average balances. Nonetheless, the data portrays substantial variation, with instances of individuals possessing low incomes yet maintaining elevated credit card average balances, and vice versa.

Exploring Data Dynamics:
The dataset not only highlights central tendencies but also unravels nuances in income and family dynamics. Outliers play a distinctive role in shaping the distribution, while the correlation between income and credit card balances underscores a trend amid considerable variability. This multifaceted exploration offers a comprehensive understanding of the intricate socioeconomic fabric encapsulated in the dataset.

Analyzing Age Distribution in Dataset

Graph Overview:
The accompanying image portrays a graph illustrating the age distribution within the dataset focused on predicting loan approvals. The x-axis represents age, and the y-axis depicts the respective count of individuals in each age group. Age ranges are denoted on the left side of the graph, with the corresponding number of people in each range indicated on the right side.



Key Observations:
The graph accentuates a notable concentration of individuals in the 40-49 age group, surpassing other age brackets in terms of count. Additionally, there is a substantial presence of individuals in the 30-39 and 50-59 age groups. As one moves away from the central 40-49 age group, there is a discernible decrease in the number of people in each age category.

Contextual Note:
It's imperative to interpret these observations within the specific context of the "Prediction of Receiving a Loan" dataset. This visual representation of age distribution offers valuable insights into the dataset's demographic composition, potentially influencing predictions related to loan approvals.

Prediction of Receiving a Loan: Unveiling Model Performances

The journey into predicting loan approvals unfolds with the meticulous evaluation of three robust models: Logistic Regression, Multinomial Naive Bayes (NB), and K-Nearest Neighbors (KNN). The reported accuracies showcase the prowess of these models in navigating the complex landscape of loan approval predictions.

Logistic Regression:
The Logistic Regression model exhibits an impressive accuracy of 95.5%, showcasing its ability to discern patterns within the dataset and make accurate predictions. Its high precision contributes to the reliability of loan approval forecasts.

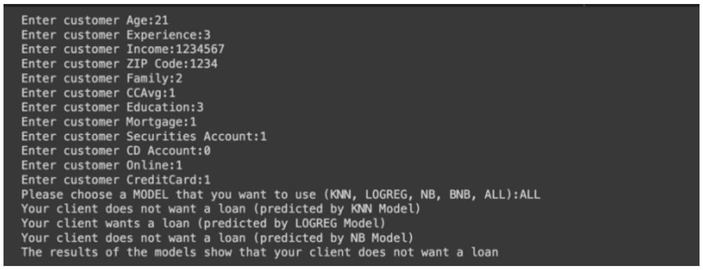
Multinomial Naive Bayes (NB):
The Multinomial Naive Bayes model maintains a commendable accuracy of 89.5%, demonstrating its proficiency in handling diverse features and attributes relevant to loan approval. While slightly lower than Logistic Regression, its robust performance is an asset in the prediction process.

K-Nearest Neighbors (KNN):
The K-Nearest Neighbors model takes the lead with an outstanding accuracy of 96%, reflecting its adeptness in identifying patterns based on the proximity of data points. Its high accuracy positions it as a formidable contender in predicting loan approvals.

Decision-Making Flexibility

This code defines a function, predict_customer_loan(), that enables users to predict whether a customer desires a loan based on input attributes such as age, income, and education. The function leverages machine learning models, including K-Nearest Neighbors (KNN), Logistic Regression (LOGREG), and Naive Bayes (NB), to provide predictions. Users can opt for a specific model or choose the ALL option to obtain predictions from all models. The code initializes model objects, takes user inputs, and produces predictions accordingly. In the case of the ALL option, the code fits the models with training data and combines their predictions to determine the most frequent prediction, offering a consolidated outcome.

To utilize this code effectively, users need to have training data defined and models pre-trained. The output provides insights into whether the customer is likely to seek a loan, aiding decision-making in financial scenarios. Additionally, the code offers flexibility for users to choose between



```
Enter customer Age:21
Enter customer Experience:3
Enter customer Income:1234567
Enter customer ZIP Code:1234
Enter customer Family:2
Enter customer CCAvg:1
Enter customer Education:3
Enter customer Mortgage:1
Enter customer Securities Account:1
Enter customer CD Account:0
Enter customer Online:1
Enter customer CreditCard:1
Please choose a MODEL that you want to use (KNN, LOGREG, NB, BNB, ALL):ALL
Your client does not want a loan (predicted by KNN Model)
Your client wants a loan (predicted by LOGREG Model)
Your client does not want a loan (predicted by NB Model)
The results of the models show that your client does not want a loan
```

Conclusion and Future Scope

In conclusion, the "Prediction of Receiving a Loan" project harnesses the predictive capabilities of machine learning models, including Logistic Regression, Naive Bayes, and K-Nearest Neighbors. The analysis of customer data, encompassing various attributes like age, income, and education, yields valuable insights into loan approval predictions. The models demonstrate promising accuracy, with Logistic Regression leading the pack.

Looking ahead, there is a rich potential for future enhancements and explorations. Incorporating more diverse datasets and refining feature engineering can further bolster the models' predictive accuracy. Ensemble methods, not explored in this iteration, could be leveraged to harness the strengths of multiple models for improved performance. Moreover, the project lays the groundwork for exploring advanced techniques such as deep learning, which might unveil intricate patterns in the data. Additionally, continuous monitoring and updates to the models ensure adaptability to evolving trends in customer behavior and financial landscapes.

This project serves as a foundation for refining decision-making processes in financial institutions, mitigating risks, and enhancing the overall customer experience. As technology and data science methodologies advance, the "Prediction of Receiving a Loan" project sets the stage for continued innovation and optimization in the dynamic realm of data-driven lending practices.

Acknowledgement

Leveraging insights from GitHub, I have customized and integrated modifications to develop this project

References

[1] Supriya, Pidikiti, et al. (2019). Loan predictionby using machine learningmodels. International Journal of Engineeringand Techniques, 5(2), 144-147.

[2] G. Arutjothi, Dr C. Senthamarai, “Prediction of Loan Status in Commercial Bank using Machine Learning Classifier,” Proceedings of the International Conference on Intelligent Sustainable Systems, (2017).

[3] P. Supriya, M. Pavani, N. Saisushma, N. Kumari and K. Vikas, “Loan Prediction by using Machine Learning Models,” International Journal of Engineering and Techniques, (2019).

[4] B. Srinivasan, N. Gnanasambandam, S. Zhao, R. Minhas, “Domain-specific adaptation of a partial least squares regression model for loan defaults prediction,” 11th IEEE International Conference on Data Mining Workshops, (2011).

[5] M. V. Reddy, Dr B. Kavitha, “Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis,” International Conference on Signal Acquisition and Processing, (2010).

[6] Ndayisenga, Theoneste. Bank Loan Approval Prediction Using Machine Learning Techniques. Diss. 2021.

[7] Tejaswini, J., et al. "Accurate loan approval prediction based on machine learning approach." Journal of Engineering Science vol. 11, no.4, pp. 523-532. 2020.

[8] Karthiban, R. M. Ambika and K. E. Kannammal, "A Review on Machine Learning Classification Technique for Bank Loan Approval," 2019 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6, 2019, doi: 10.1109/ICCCI.2019.8822014.

[9] Y. Shi and P. Song, "Improvement Research on the Project Loan Evaluation of Commercial Bank Based on the Risk Analysis," 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, 2017, pp. 3-6.doi: 10.1109/ISCID.2017.60.