LOAN STATUS PREDICTION MODEL

Business Understanding.

In the dynamic realm of financial services, the ability to swiftly and accurately process loan applications while delivering exceptional customer service stands as a fundamental competitive edge.For a long time customers have wasted a lot of time on queues waiting for loan approvals only to end up disappointed.The inception of our Loan Prediction Model integrated with Chatbot technology aims to redefine the customer journey in loan processing by enhancing both the efficiency and the quality of service delivered. This innovative solution is poised to transform how financial institutions such as banks,saccos,shylocks interact with their clients, making loan accessibility quicker and more user-friendly.

The challenges faced by traditional loan approval methods are multifaceted. Manual review processes are often time-consuming and susceptible to human error, leading to inconsistencies and potential biases. Furthermore, the inability to effectively analyze vast amounts of data can hinder the identification of key risk factors, resulting in suboptimal decisions that may contribute to higher default rates.

A data-driven approach leveraging machine learning techniques presents a transformative solution. By developing a predictive model capable of accurately classifying loan approval outcomes, financial institutions can automate decision-making processes, significantly reducing the time and resources required for manual reviews. This automated system will not only enhance operational efficiency but also ensure consistent and objective evaluations, minimizing the risk of defaults.

The overarching goal is to streamline the loan approval workflow, minimizing the time required for decision-making while simultaneously reducing the potential for defaults. Successful implementation of this predictive model will position the financial institution as industry leaders, offering a seamless and reliable lending experience to their valued customers.This strategic initiative will not only streamline operations and mitigate risks but also cement the institution's reputation for innovation and customer-centric services, fostering trust and loyalty among its clientele.

Research Question.

- What is the eligibility criteria for getting a loan status as qualified
- Which age group are mostly involved in loan borrowing
- How does account balance, loan at the time of checking loan status and housing loan affect loan status
- Which job , education level and marital status influence loan status?

Problem Statement.

To assist financial institutions in automating and optimizing their loan approval processes.

By leveraging historical data, we aim to build a reliable model that predicts whether a loan application is likely to be approved or rejected.

Objectives.

Project Objective

Why automate a Loan approval process?

- To automate the Loan Approval Process: By automating the evaluation of loan applications using advanced machine learning techniques, we aim to significantly decrease decision times and increase throughput without compromising accuracy.
- To boost customer engagement: Through a ML chatbot, customers can receive guidance on their loan applicationsto ensure a seamless and informative experience at every interaction point.
- Risk Minimization: The model will help minimize the risk of approving loans that may default.
- Enhances Operational Efficiency: By reducing manual intervention in the loan approval process, our solution aims to cut operational costs and redirect human resources towards more strategic, high-value tasks.

Data Understanding.

We sourced the dataset from github dataset repository. The dataset encompasses a wealth of information about individuals and their interactions with a financial institution. It includes customer details, financial status indicators, and records of past marketing campaigns. Notably, the dataset contains a target variable 'y' that signifies whether an individual had his/her loan approved.In addition to the existing features, the data understanding process may reveal opportunities for feature engineering. This involves creating new features or transforming existing ones to better capture the underlying patterns and relationships within the data

Data Description.

The dataset had 18 columns and 4521 rows. The columns in the dataset included the following

- **age:** The age of the individual targeted in the campaign.
- **job:** The job role of the individual (e.g., manager, blue-collar, technician, etc.).
- **marital:** The marital status of the individual (single, married, divorced).
- **education:** The highest level of education achieved by the individual (secondary, tertiary).
- **default:** Indicates whether the individual has credit in default (yes or no).

- **balance:**The account balance of the individual
- **housing:** Indicates whether the individual has a housing loan (yes or no).
- **loan:** Indicates whether the individual has a personal loan (yes or no).
- **contact:** The type of communication used to contact the individual (cellular, uknown).
- **day:** The day of the month on which the last contact was made.
- **month:** The month during which the last contact was made
- **duration:** The duration of the last call, in seconds.
- **campaign:** The number of contacts performed during this campaign for this client (includes last contact).
- **pdays:** The number of days that passed since the client was last contacted from a previous campaign (999 means client was not previously contacted).
- **previous:** The number of contacts performed before this campaign for this client.
- **poutcome:** The outcome of the previous marketing campaign (e.g., success, failure, nonexistent).
- **y:** The target variable indicating whether the campaign was successful with this client (yes or no).

Data Preparation.

Loading the data.

After that, the datasets was added to the Jupyter Notebook, where it was previewed for a better understanding of their columns and the relationships that exist between them.

Cleaning data.

From the data frame data cleaning was performed. For easier comprehension, the following columns **LP,y,balance** were **renamed** into **Loan purpose,acc balance** and **loan status** columns. Subsequently, the dataframe was scrutinized to check for **missing values** and there were no missing values and no duplicates

Feature Engineering.

Feature Engineering work was done to create a new column **loan access and repayment amount.** The loan access was created by multiplying acc balance time 10 and repayment amount was calculated by multiplying the loan access* an interest rate of 12% .Then the **loan access** column was converted from string datatype to integer datatype and finally

EDA

Performing a comprehensive exploratory data analysis (EDA) is crucial for visualizing our column features for easier understandability. The analysis focused on key dataset features, such as the age distribution,marital status job titles and education. Visualizations like histograms,bar graphs, box plots, and violin plots provided insights. Bar graphs revealed distributions of loan status, marital status,job counts, and education level counts. Box plots depicted acc balance.

Modeling

The project started with baseline model with logistics regression being chosen . It was noticed that the baseline model performed well but because there was a class imbalance with the loan status yes being the minority compared to the no status.

We then moved to decision tree model. In this model we addressed the class balance using SMOTE and the model accuracy was 75%. We then tried using the oversampler to adress the class imbalance and the accuracy went to 80%.After trying cross validation on the decision tree we got an average accuracy of 79%.We then moved to **support vector machine** and the the best cross validation score from tuning the hyperparameters wass 74% accuracy using SMOTE to address class imbalance using random over sampler reduces the accuracy to 70%.The best accuracy score from the grid search was 70.82% which is a drop from the accuracy registered in Decision trees. Addressing the class imbalance using random over sampler reduced the grid search accuracy t0 64%

We then moved to **random forest**, the accuracy was 79.23%. The model correctly predicted the outcome 79.23% of the time across both classes. The grid search identified max_depth of 30 and n_estimators of 200 as the best parameters for the classifier and a cross-Validation accuracy of (79.76%).The performance on the test set was 78.78%.

We then tries two gradient boosting model the adaboost and XGB boost model.For the **adaboost** the adaboost model gave an accuracy of 69.72% which was by far the lowest accurcay recorded with a precision accuracy of 91% and a recall of 73%. Adressing class imbalance with random over sampler gave an ccuracy: 0.6519% which was very low.The XGB boost model accuracy was 82%.The best cross validation score accuracy from tuning XGB boost was 83% with a test accuracy of 84.20% with the best parameters being {'classifier__learning_rate': 0.1, 'classifier__max_depth': 6, 'classifier__min_child_weight': 1, 'classifier__n_estimators': 200, 'classifier__subsample': 0.7}.This was by far the best score recorded from all the models tuned

We also tried the **neural network** and the test accuracy of the neural network model was 89.17% a great improvement from the XGB model which was 84% making it the best model in terms of accuracy so far.The naive bayes accuracy was 81.42% suggested that the Naive Bayes model was able to correctly predict the outcome for approximately 81.42% of the training set.The testing accuracy was 82.54% . Finally we tried the **catboost model** .The catboost classifier had an accuracy of 88.41 on the training data and an accuracy of 89.17 on the testing accuracy which was a good accuracy the best score for this model after tuning the parameters was 88.3% which was slightly higher than the neural network accuracy score of 88.19%.This model was well-suited for categorical data, hence the name "CatBoost" for "Category" and "Boosting." its ability to handle categorical variables directly. Unlike other machine learning models that require extensive preprocessing to convert categorical variables into numerical formats, CatBoost can process these features as they are

User interaction

For the user interaction we created a chatbot on our deployment platform where the user can interact and ask questions.The user will be prompted to enter his/her details and the information displayed to him/her. The detailed information about the user also displays the amount of loan he/she is able to access and also get the repayment amount required.The user can ask for help in terms of checking his/her loan status and will be directed accordingly on how to apply for a loan incase he was qualified to continue for a loan request.
On the deployment platform the user is able to calculate his/her preferred repayment amount to be able to plan accordingly.

Conclusions

In conclusion, this project has successfully achieved its primary objective of developing a comprehensive and interactive loan status prediction system. This system not only assesses loan eligibility based on a variety of user-inputted financial and personal details but also leverages a sophisticated machine learning model to ensure accuracy and reliability in predictions. The integration of the CatBoost classifier allows the system to handle complex datasets with a mix of categorical and numerical features effectively, providing users with precise loan status assessments.Throughout this project, several key objectives were accomplished. We designed and implemented a user-friendly interface using Streamlit, which simplifies the process for users to input their data and receive immediate loan status predictions. Additionally, we performed an extensive analysis of the factors that significantly influence loan approval decisions. This deep understanding was critical in fine-tuning our predictive model to ensure that it delivers relevant and trustworthy predictions to the users.Furthermore, we enhanced user engagement by incorporating a responsive chatbot within the application. This chatbot assists users by answering common queries related to loan applications and navigating through the application features, thus enriching the overall user experience. By offering instant responses and guidance, the chatbot ensures that users feel supported throughout their interaction with the platform.
In summary, this project has taken a holistic approach to creating a robust loan status prediction system that not only meets its specified objectives but also significantly improves the loan application process. Users benefit from a system that not only predicts loan eligibility with high accuracy but also provides insights into the factors affecting their loan status, all within a supportive and interactive environment. This project has set a solid foundation for future enhancements, including the integration of more dynamic features and the adoption of more advanced machine learning algorithms to further enhance prediction accuracy and user experience.

Recommendations

1. **Continuous Model Monitoring and Updating**: It is recommended to implement a continuous monitoring and updating process for the predictive model to ensure its reliability and relevance as new data becomes available or market conditions change.
2. **Explainable AI**: Incorporating explainable AI techniques can enhance the interpretability and transparency of the model's predictions, facilitating trust and understanding among stakeholders and regulatory bodies.
3. **Feature Engineering and Ensemble Methods**: Further exploration of advanced feature engineering techniques and ensemble methods could potentially yield performance improvements and capture more complex patterns and relationships in the data.
4. **Responsible AI Practices**: As the model is deployed in a financial context, it is crucial to adhere to responsible AI practices, ensuring fairness, accountability, and compliance with relevant regulations and ethical guidelines.
5. **Chatbot Expansion**: Expanding the scope of the chatbot to provide more comprehensive financial advisory services beyond loan applications could enhance customer experience and foster long-term loyalty.
6. **Collaboration and Knowledge Sharing**: Fostering collaboration and knowledge sharing among data scientists, domain experts, and stakeholders can facilitate the integration of diverse perspectives, leading to more robust and impactful solutions.

Future Improvement Ideas

- **Integration of More Data Sources**:Expanding the dataset to include more variables such as credit score, employment history, and existing financial obligations could improve the model's predictive accuracy.
- **Advanced Machine Learning Techniques**:Exploring more sophisticated machine learning algorithms and ensemble methods could enhance the predictive performance.
- **Personalized User Experience**:Developing a more personalized dashboard that tracks and displays individual loan application history, suggested actions to improve loan eligibility, and personalized financial advice could enhance user engagement and satisfaction.
- **Automated Feature Engineering:**Implementing automated feature engineering techniques can help in discovering more meaningful patterns and relationships in the data, which could improve model performance without manual intervention.
- **User Feedback Loop:**Establishing a mechanism to collect user feedback on the application process and the accuracy of the loan predictions can provide insights into the system’s performance and areas for improvement.
- **Integration with Financial Institutions:**Partnering with banks and other lending institutions could allow for seamless data sharing and validation, improving the accuracy of the predictions and possibly integrating the system directly into the loan application process of these institutions.

Deployment

Following rigorous testing and evaluation, we have selected the best-performing model to underpin our loan status prediction system. To ensure the longevity and reusability of the model, we will serialize it using the 'pickle' method. This process will capture the model's current state, allowing it to be efficiently reused in future applications. The model will be integrated into a user-friendly web interface designed which is streamlit that is easier to incorporate with the chatbot.The model will be hosted by the streamlit servers which makes it easier.This interface will feature a dynamic loan status prediction tool that provides users with real-time feedback on their loan eligibility based on their financial data and other relevant parameters.Additionally, we will explore the integration of this system into existing banking software systems, allowing financial institutions to provide instant loan status updates to their customers.This comprehensive deployment approach aims to make the loan status prediction tool accessible, convenient, and practical for a wide audience, thereby enhancing financial decision-making for individuals and lenders alike.