Aakash Ojha

Data Practicum

Final Project

**Data Practicum**

**Executive Summary**

**Defaulter**

1. **Introduction**

Lending institutions must efficiently assess loan applications in the fast-paced financial environment of today to reduce the risk of defaults. Loan default happens when borrowers don't fulfill their repayment commitments, which causes financial losses for lenders and could influence the overall economy. By making precise predictions of the risks of loan default, machine learning techniques provide a potent answer to this problem.

The goal of this research is to create a prediction model that will help lending institutions evaluate loan applications more intelligently. We can build a trustworthy model that recognizes patterns and risk factors related to loan defaults by evaluating historical loan data and utilizing cutting-edge machine learning methods. Lenders may find this predictive model to be a useful tool for streamlining the loan approval process, lowering default rates, and ultimately improving portfolio performance.

We will examine and preprocess a large dataset of loan applications during this project, which includes a variety of information like applicant demographics, credit ratings, loan characteristics, and previous financial behaviors. To find significant insights and illustrate trends that can direct feature engineering and selection, we will go into exploratory data analysis (EDA). Then, we'll use machine learning strategies to develop, train, and improve a predictive model, ranging from simple algorithms to complicated ensemble methods.

Our goal is to build a machine learning model that performs well while still being easy to understand. We will explore feature importance analysis, which will reveal the elements that have a substantial impact on forecasts of loan default. By doing this, we can offer lending institutions with useful information that they can use to change their lending practices and understand the factors that influence the model's projections. The app name I will be using is defaulter this app will be provided specifically to financial institution like bank and other lending companies. Using this app banks will be able to find who will be the people who are not going to be paying their loan on time or will miss a lot of installments. This will help with them to make sure the customer base they have are good and they will be paying their loan on time. For the practicum I wanted to use machine learning model which will predict who will be paying their loan on the time and who will not. The three major data science model I'm using or logistic regression, decision tree and naïve bayes. I will be comparing all these models to see which algorithm has better accuracy and other statistical data as well. The reason I have decided to change my topic is because the data set, I used previously was insufficient and I was not able to showcase my eda as well as the visualization skills in the other data set since the number of data available was not as good. So let me explain how I did my EDA process as well as the libraries which I used for the manipulation process as well as the visualization process.

Since there are 34 different columns, so here is an explanation for them:

1. **ID:** A unique identifier for each loan application.
2. **year:** The year in which the loan application was processed.
3. **loan\_limit:** Information about the loan limit for the application.
4. **Gender:** Gender of the applicant.
5. **approv\_in\_adv:** Whether the loan was approved in advance.
6. **loan\_type:** Type of the loan.
7. **loan\_purpose:** Purpose of the loan.
8. **Credit\_Worthiness:** Assessment of the applicant's creditworthiness.
9. **open\_credit:** Information about open credit lines.
10. **business\_or\_commercial:** Whether the loan is for a business or commercial purpose.
11. **loan\_amount:** The amount of the loan applied for.
12. **rate\_of\_interest:** The rate of interest for the loan.
13. **Interest\_rate\_spread:** The spread between the loan's interest rate and a benchmark rate.
14. **Upfront\_charges:** Any upfront charges associated with the loan.
15. **term:** The term (duration) of the loan.
16. **Neg\_ammortization:** Whether the loan has negative amortization.
17. **interest\_only:** Whether the loan has interest-only payments.
18. **lump\_sum\_payment:** Whether there's a lump-sum payment requirement.
19. **property\_value:** The value of the property being financed.
20. **construction\_type:** Type of construction.
21. **occupancy\_type:** Type of occupancy.
22. **Secured\_by:** What the loan is secured by.
23. **total\_units:** The total number of units in the property.
24. **income:** The applicant's income.
25. **credit\_type:** Type of credit.
26. **Credit\_Score:** Credit score of the applicant.
27. **co-applicant\_credit\_type:** Credit type of the co-applicant.
28. **age:** Age of the applicant.
29. **submission\_of\_application:** How the application was submitted.
30. **LTV:** Loan-to-value ratio.
31. **Region:** Geographic region.
32. **Security\_Type:** Type of security for the loan.
33. **Status:** The status of the loan.
34. **dtir1:** Debt-to-income ratio.
35. Problem Statement

The main goal of this project is to create and assess predictive models that can precisely estimate the probability of loan defaults based on a wide range of input data. This study aims to improve risk assessment in the lending sector by utilizing the capabilities of data science techniques including logistic regression, naive Bayes, and decision tree algorithms. Our goal is to give financial institutions trustworthy tools for proactive decision-making through the analysis of historical loan data that includes a variety of borrower attributes and economic indicators, allowing them to more efficiently allocate resources and reduce potential losses from loan defaults.

**Relevance in the Prediction of Loan Default Context:**

In the financial industry, where lending institutions are frequently exposed to significant risks because of borrower uncertainty and economic changes, loan default prediction is of the utmost importance. These institutions are better equipped to make educated judgments, allocate resources wisely, and create specialized risk management methods when they can accurately forecast if a borrower is going to fail on their loan. The following essential factors will help you comprehend the value of this project:

Risk management and loss prevention: Lenders may suffer significant financial losses because of loan defaults. An early warning system can be provided by precise prediction models, allowing lenders to proactively manage risks and reduce possible losses.

Resource Allocation: Lending institutions can distribute their resources more effectively if they are aware of high-risk borrowers in advance. This entails changing loan terms, interest rates, or even turning down high-risk loan applications.

Better Decision-Making: Lenders may decide on loan approvals and terms with greater knowledge thanks to data-driven insights from predictive models. This lessens the likelihood of lending to clients who might have repayment issues.

Prediction models help with informed portfolio management by helping to keep a balanced and varied loan portfolio. This diversity lessens the risks brought on by economic downturns or vulnerabilities in particular industries.

Governing Compliance: Various regulatory agencies impose strict rules on risk assessment on financial institutions. Effective prediction models assist institutions in upholding these criteria and proving that they exercise due diligence when making loans.

Enhanced Customer Relationships: Lenders can keep good ties with borrowers by preventing defaults through focused risk assessment. Increased client satisfaction and repeat business may result from this.

Economic Stability: By lessening the detrimental rippling effects that default waves can have on the financial system, preventing loan defaults helps to maintain general economic stability.

The project's relevance, then, resides in its ability to fundamentally alter how lending institutions evaluate and manage risks. The use of data science approaches to predict loan defaults can lead to more informed lending decisions, safeguard financial stability, and strengthen bonds between lenders and borrowers. The project is in line with the larger objective of creating sustainable lending practices and economic growth by addressing this important part of the financial sector.

1. Data Collection and Preprocessing

**Data Collection Process:**

A carefully chosen dataset for this project was obtained from Kaggle, a well-liked website for datasets and data science competitions. The dataset, termed "[Loan Default Dataset](https://www.kaggle.com/datasets/yasserh/loan-default-dataset)," is intended to aid in the analysis and prediction of loan defaults based on various borrower and loan variables. The dataset is freely accessible and can be found at the following URL: Kaggle's Loan Default Dataset.

The dataset consists of records for past loan applications, each of which has a collection of attributes that reveal information about the borrower's profile, the terms of the loan, and economic indicators. The creator of the dataset most likely obtained this information by combining data aggregation from lending-related sources such as credit bureaus, financial institutions, and potentially other sources.

**Preprocessing Summary:**

The project's initial phases required thorough data preprocessing, a crucial step carried out with the help of adaptable tools like pandas, Matplotlib, and Seaborn. The foundation for further analysis and model creation was laid during this crucial phase. The primary preprocessing tasks that were carried out can be summarized as follows:

Exploratory Data Analysis (EDA): The dataset was initially loaded using pandas, allowing for a thorough examination of both its composition and structure. The EDA provided a rudimentary understanding of the variables and their interactions, acting as a steppingstone.

Visual insights: Using seaborn and matplotlib, the data was brought to life through perceptive visuals. Visual representations were made of the distribution patterns, potential abnormalities, and correlations. I used a heatmap to see the correlation between the columns, this helped me to decide which columns can be essential to make the prediction.

Treatment of Null Values: Since maintaining the integrity of the data is crucial, an assessment of the missing values was made. There were so many null entries in some columns that it was purposefully decided to remove all rows with missing data. This wise decision protected the dataset's analytical fidelity.

Determining the Significance of Features: The visualization toolbox made it easier to find important features. These characteristics were chosen for more thorough examination and modeling because of their impact on the prediction of loan default.

Age Categorization Transformation: A code-based transformation was used to address the categorical classification of age, which resulted in a numerical representation. This translation gave the age feature numerical continuity, making it suitable for incorporation into predictive models.

Utilizing Selective Attributes: Due to the dataset's diverse makeup, a discriminating strategy was used, concentrating on attributes specifically relevant to the prediction of loan default. The wise choice of attributes reduced unnecessary complexity and streamlined following procedures.

Preparing Categorical Variables: Due to the complexity of categorical variables, a change in encoding was required to make them compatible with machine learning techniques. To give categorical qualities a numerical core that would allow for easy incorporation into modeling processes, one-hot encoding was used.

Preparing the Train-Test Split: The dataset was split into subgroups for training and testing before the modeling process. This separation made it easier to evaluate the performance of the model, validate it, and gauge its potential to generalize.

In conclusion, the dataset was prepared for predictive modeling through the rigorous use of data preprocessing, in conjunction with insights gained from visualization and analytical discernment. This preliminary stage serves as the foundation for the upcoming loan default prediction models by carefully handling missing values, feature selection, transformation, and encoding.

1. **Exploratory Data Analysis (EDA): Recognizing the dynamics of loan default.**

Exploratory data analysis (EDA) was rigorously conducted to shed light on the complex dynamics of loan default prediction. This was done by utilizing cutting-edge visualization methods and careful correlation analyses. The goal of this project was to identify hidden patterns, gain knowledge, and provide guidance for the tactical creation of prediction models for loan default scenarios. The following main conclusions have been drawn from the resulting observations:

Stratification of Loan Amounts and Defaulters: Careful visualization revealed a pronounced split in loan amounts between responsible payers and defaulters. Notably, loan defaulters tended to favor smaller loan values, raising the possibility of a link between lower loan values and a higher propensity for default.

Geographical Implications and Default Trends: Geographical segmentation produced enlightening information about loan default trends. The "North" and "South" regions showed high loan applicant densities. The prevalence of loan defaults was mirrored by this demographic density, indicating a complex link between regional distribution and default occurrences.

Influence of Pre-Approval on Default Results:

Looking closely at the variable "loan approved in advance," an unusual pattern became apparent. When compared to borrowers who had loans pre-approved, those who did not showed a noticeably greater rate of defaults. The possible influence of pre-approval procedures on loan repayment behavior is highlighted by this.

Heatmap-Based Discernment of the Correlation Matrix: The complex interplay between the variables was made clear by a heatmap-based interpretation of the correlation matrix. The result underlined how important it is to recognize and pick relevant elements for strong modeling. We considered features with poor correlations to the target column to be irrelevant for prediction tasks.

Age Dynamics and Default Trends: A careful analysis of the age demographics revealed that people between the ages of 55 and 74 represent a critical age group. In contrast to other age cohorts in the dataset, this cohort showed a higher probability for loan defaults. Recognizing this different age-related pattern helps us to understand default behavior in a more complex way. Influence of Loan Types on Defaults:

Thorough analysis of loan types unveiled a pronounced prevalence of "type one" loans within the dataset. This dominance was observed to eclipse the cumulative occurrence of other loan types. Remarkably, the higher prevalence of "type one" loans corresponded to an elevated incidence of defaults within this category, accentuating the intricate interplay between loan type and default occurrences.

Kernel Density Insights into Loan Amounts:

The kernel density estimation facilitated a granular understanding of loan amount distributions. Of note, the highest density peak centered around the 0.3% range, signifying a pivotal range with the highest frequency of loan amount occurrences. This insight empowers data-driven decision-making in loan approval strategies.

In conclusion, the comprehensive EDA undertaken has engendered a nuanced comprehension of loan default dynamics. The synthesis of insights spanning regional, demographic, and categorical facets imbues the predictive modeling process with an informed foundation. The culmination of these findings not only enriches our understanding of loan default behavior but also facilitates the construction of predictive models that can anticipate and mitigate the risk of loan defaults with heightened accuracy.

Below are some of the graphs I used for the above-mentioned analysis:

A graph with green and orange bars

Description automatically generatedA blue and orange rectangular bars

Description automatically generatedA graph with blue squares

Description automatically generatedA screenshot of a graph

Description automatically generatedA graph of a graph

Description automatically generatedA graph of different types of data

Description automatically generatedA graph of a graph of a loan amount

Description automatically generated

1. **Methodology:**

For the methodology I am referencing the information which was given on the Kaggle website:

**‘Description:**

Banks earn a major revenue from lending loans. But it is often associated with risk. The borrowers may default on the loan. To mitigate this issue, the banks have decided to use Machine Learning to overcome this issue. They have collected past data on the loan borrowers & would like you to develop a strong ML Model to classify if any new borrower is likely to default or not.

The dataset is enormous & consists of multiple deterministic factors like borrower’s income, gender, loan purpose etc. The dataset is subject to strong multicollinearity & empty values. Can you overcome these factors & build a strong classifier to predict defaulters?

**Acknowledgements:**

This dataset has been referred from Kaggle.

**Objective:**

* Understand the Dataset & cleanup (if required).
* Build classification model to predict weather the loan borrower will default or not.
* Also fine-tune the hyperparameters & compare the evaluation metrics of various classification algorithms.’ (Kaggle Website Information)

And I will be using the following method in my research:

1. Logistic Regression: Based on input features, logistic regression is a classification technique that forecasts binary outcomes. It divides cases into one of two classes by comparing the probability score to a threshold and then applies a logistic function to translate a linear combination of features into a probability score.

2. Naive Bayes: Based on the conditional probabilities of a feature's attributes, Naive Bayes determines the likelihood that an instance belongs to a class. It makes text categorization and spam detection effective since it assumes feature independence.

3. A decision tree is a flexible model for regression and classification. It produces a tree-like structure by repeatedly dividing the data according to feature criteria. By moving from the root to a leaf node, which represents a class prediction or a regression value, it makes decisions.

Each algorithm offers unique advantages: Naive Bayes is effective for text, Logistic Regression offers interpretability, and Decision Trees may capture intricate relationships. The decision is based on the data's properties and modeling objectives.

1. **Model Training:**

In the initial stages of the study, it was crucial to separate the columns, where the variable x was carefully selected to include key characteristics crucial for the prediction of the target variable, particularly the "Status" column. Following this pre-processing step, the well-known train-test split technique was used to smoothly divide the x and y variables into separate training and testing datasets. A fair distribution was maintained in this partitioning strategy, with a careful allocation of 25% of the dataset going to the testing subset and the remaining 75% going to the training subset. The result of this project took the form of a dataset that was ready for later modeling.

The next step included integrating the Sklearn library's logistic regression module, which serves as the basis for predictive modeling. A classifier was successfully instantiated, creating the framework for the use of predictive analytics. This classifier made it easier to anticipate the "Status" variable, which is essential to the process of predicting loan defaults. Several core indicators that are essential for thorough model evaluation were carefully used. These included the accuracy score, precision score and recall score functions from the Sklearn package. They also included F1-score and ROC-AUC score. The remaining steps of the analysis played out similarly for the alternate models being thought about. The adaption included the selective importation of pertinent models from the Sklearn library in alignment with the desired predictive algorithms while retaining methodological homogeneity. By methodically applying the metrics suite to all models, performance benchmarks could be consistently derived, providing a solid foundation for later comparison evaluations.

1. **Model Evaluation:**

Model 1 – Logistic Regression

Model 2 – Decision Tree

Model 3 – Naïve Bayes

A clear trend can be seen after carefully examining the graphical representations below: All the model provides equal accuracy which means that all the three model would predict the same.

During the evaluation of the logistic regression's performance indicators, an unusual finding emerged. The precision, recall, and F1 score all showed a manifestation of 0. This result, however surprising, highlighted a significant difference from the other two models that showed data from all three of the statistical studies. Contrarily, a closer look at the ROC curves revealed subtle differences between Model 1 and Model 3, which may indicate a more complex interaction.

When all factors are considered, the decision tree model is supported by most of the data. The decision tree approach emerges as the best prediction model for this dataset, while admitting minor variations. This choice is supported by the model's outstanding performance across a range of evaluation measures and is consistent with the broader goal of achieving reliable and accurate loan default predictions.

A comparison of evaluation metrics

Description automatically generated

A screen shot of a computer

Description automatically generatedA screen shot of a computer

Description automatically generatedA screen shot of a computer

Description automatically generated

1. Limitations and Future Work:

In hindsight, the dataset had some issues, mostly because there were so many missing values. Theoretically, increased predictive power could be attributed to a bigger volume of complete data. A further problem was the presence of multiple irrelevant columns, which made it difficult to use them to build a more accurate prediction model because they had no bearing on the target variable. The desire for more relevant raw data highlights the possibility of more accurate predictions if the dataset were inherently more matched with the current prediction task. The effort for thorough model assessment is commended, which encourages further research into additional modeling methodologies beyond the current three. Although the models produced similar accuracy, investigating alternative algorithms could offer insightful details on the performance of prediction under various techniques.

The creation of a finance-focused application, which is envisioned as a useful tool for commercial deployment, is a practical next step. The goal is to provide businesses with a resource that successfully uses data-driven insights for the best decision-making by leveraging my background in finance. A strong data collection system is also in the works, which will be crucial for gathering information for upcoming research projects. The ethical necessity of collecting data is adequately acknowledged, leading to the creation of a transparent framework that clarifies data usage and safeguards and promotes user confidence.

Together, these future endeavors demonstrate an emerging path toward improved prediction models, useful applications, and thoughtful data use. The culmination of these strategic objectives denotes the path towards an advanced and morally upstanding application of data science in the financial industry.

1. Summary and Findings:

Data Preprocessing Insights: The dataset was cautiously evaluated through rigorous preprocessing using Pandas, Matplotlib, and Seaborn, revealing the importance of handling missing values and choosing pertinent characteristics. Visualizations showed how many factors affected the likelihood of a loan default.

Wealth Inequality and Default Rates: A closer look at loan amounts revealed that defaulters frequently requested smaller loans than those who made timely payments. Additionally, a correlation between densely inhabited areas and a higher proportion of default situations was revealed through regional research.

Pre-Approval and Default Rates: Non-pre-approved applicants showed a larger propensity to default, indicating the importance of pre-approval in loan defaults. This demonstrates how crucial pre-approval procedures are in influencing borrower behavior.

The relationship between age and loan default tendencies showed a clear pattern, with people in the 55–74 age range being more likely than people in other age groups to fail on their loans. This demographic information helps to elucidate default processes in a complex way.

Model Performance and Comparison: The decision tree algorithm's greater accuracy in comparison to other models was demonstrated via model evaluation. Unexpected findings from logistic regression were shown, yet ROC analysis revealed minute variations in model performance.

Future Direction and Application: The project's conclusion highlights the value of looking into other models and methodologies. The long-term goal is to create a financial application that combines subject knowledge with moral data collection techniques.

Together, these important findings demonstrate the complex dynamics of loan default prediction, highlighting the significance of preprocessing, model evaluation, and the application of insights for practical financial decision-making.

For financial institutions to efficiently manage risk, utilize resources sensibly, and make educated lending decisions, accurate loan default prediction is essential. It aids in asset preservation, capital utilization optimization, compliance with regulations, upkeep of client connections, and general financial stability. In essence, accurate default prediction improves risk management strategies and decision-making, preserving the institution's financial stability and reputation.