



Comprehensive Analysis of Banking Loan Data and Predictive Model Evaluation

College of Professional Studies, Northeastern University

ALY 6040: Data Mining Applications

Professor. Kasun Samarasinghe

February 14, 2024

Group 6 Members:

Rohith Krishnamurthy

Likith Garimella

Maheswar Veerlanka

Likith Sai Challa

Contents

Introduction	3
Data Exploration and Cleaning	3
Methodology	10
Data Preparation	11
Model Evaluation	11
Loan Status	11
Interest Rate	12
Findings and Insights	13
Web Application:	13
Recommendations.....	15
Conclusion.....	15
References.....	16

Introduction

In a landscape where data-driven decisions are paramount, this dataset serves as a crucial resource for understanding the dynamics of the lending industry. With a blend of numerical and categorical data, it provides a comprehensive overview of loan amounts, interest rates, and other pivotal features. The goal of this analysis is to transform this extensive data into actionable insights, aiding in risk management, financial decision-making, and a nuanced understanding of borrower behaviour.

The objective is to discern pivotal driver variables impacting loan approval decisions in a consumer finance company by extensively analysing a loan request dataset in Python. Formulate strategies for denying loans, reducing loan amounts, and setting higher interest rates for risky applicants. The analysis aims to enhance decision-making processes, ensuring alignment with the dataset's inherent information for a nuanced understanding of consumer finance dynamics and risk assessment in the lending domain.

Employing a combination of statistical analysis and data visualization techniques, we will dissect the dataset to reveal underlying trends, identify patterns, and highlight anomalies in loan disbursements. This report is not just an examination of data points but a narrative that reflects the financial journeys of thousands of borrowers, as well as the risk and return considerations for lenders.

Data Exploration and Cleaning

The "Banking Loan Case Study Data" from Kaggle presents a comprehensive dataset for in-depth analysis, comprising 39,717 rows and 111 columns. This dataset offers a robust foundation for examining various facets of loan issuance, covering a diverse array of loan-related attributes. Exploring the dataset provides an opportunity to unravel critical insights into the complexities of loan distribution, borrower profiles, and overarching lending trends within the banking domain.

The dataset includes an extensive list of variables, and some specific variables of interest, along with their data types, are highlighted below:

1. **Loan Amount (Numeric):** The principal amount requested by the borrower.
2. **Interest Rate (Numeric):** The annual interest rate for the loan.
3. **Loan Term (Categorical):** The duration of the loan (e.g., 36 months, 60 months).
4. **Borrower's Credit Score (Numeric):** The credit score of the loan applicant.
5. **Employment Status (Categorical):** The current employment status of the borrower.
6. **Loan Purpose (Categorical):** The intended use of the loan funds (e.g., debt consolidation, home improvement).
7. **Home Ownership (Categorical):** The borrower's housing situation (e.g., own, mortgage, rent).

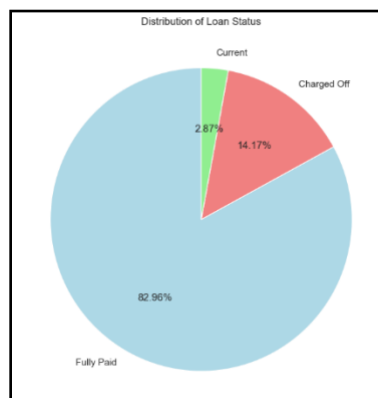
8. **Loan Status (Categorical):** The outcome of the loan application (e.g., fully paid, charged off).

Our initial interest in this dataset stems from its substantial size and the diverse set of variables, providing an opportunity to conduct a nuanced analysis of factors influencing loan approval, repayment patterns, and the overall landscape of the lending process. We aim to uncover patterns, relations, and potential predictors that contribute to informed decision-making in the context of loan management. Additionally, we are interested in identifying any emerging trends or insights that can enhance our understanding of borrower behaviour and guide strategic decisions within the banking sector.

In the data cleaning process, the initial steps involved identifying and handling missing values in the loan dataset. We begin by calculating the count and percentage of null values in each column, and subsequently, columns with all null values are removed. This ensured a more focused analysis of relevant features. Further refinement was carried out by eliminating columns with more than 50% null values, effectively reducing dimensionality, and focusing on attributes with substantial data coverage. The resulting dataset was then inspected for duplicate records based on the 'id' column, confirming the absence of duplicate entries.

Following this, additional features are engineered to enhance the dataset's utility. A new 'loan period' column is created by extracting numerical values from the 'term' column, providing insights into loan duration. The 'zip_code' column is transformed into a numeric 'zip_code_num' column after removing non-numeric characters. Redundant columns, including 'emp_title', 'url', 'desc', 'title', 'zip_code', and 'term', are dropped to streamline the dataset. The 'earliest_cr_line' column is also removed, contributing to a more refined dataset for subsequent analysis. The resulting dataset is thoroughly cleaned, eliminating redundancies, and enhancing its suitability for in-depth analysis and modelling in the context of consumer finance, and we end up with a data frame consisting of 47 variables.

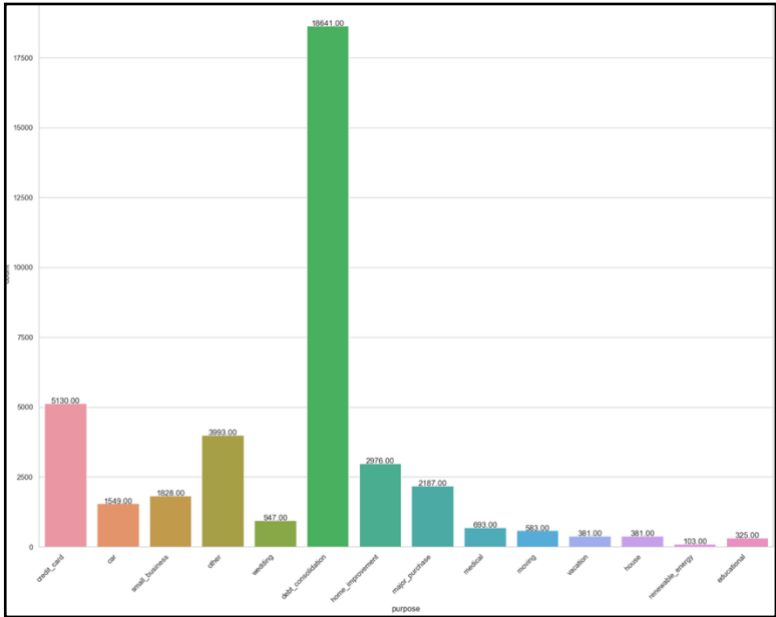
Fig1: Loan Status Proportions in the Dataset



The pie chart from **Fig1** illustrates the distribution of loan statuses within the dataset. A vast majority of loans, 82.96%, are fully paid, indicating a high rate of successful repayments. Charged-off loans, which are unlikely to be repaid, make up 14.17%, highlighting a risk of loss. Only a small fraction, 2.87%, are currently active and being repaid on time. This distribution

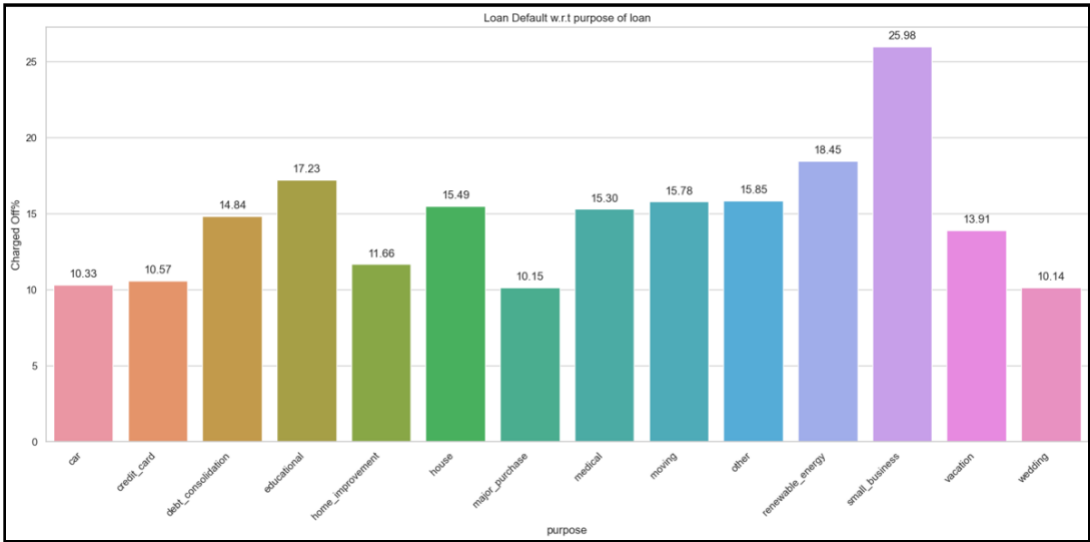
suggests that while most borrowers are reliable, a focus on the factors leading to charge-offs could enhance lending risk assessments.

Fig2: Loan Distribution by Purpose in the Dataset



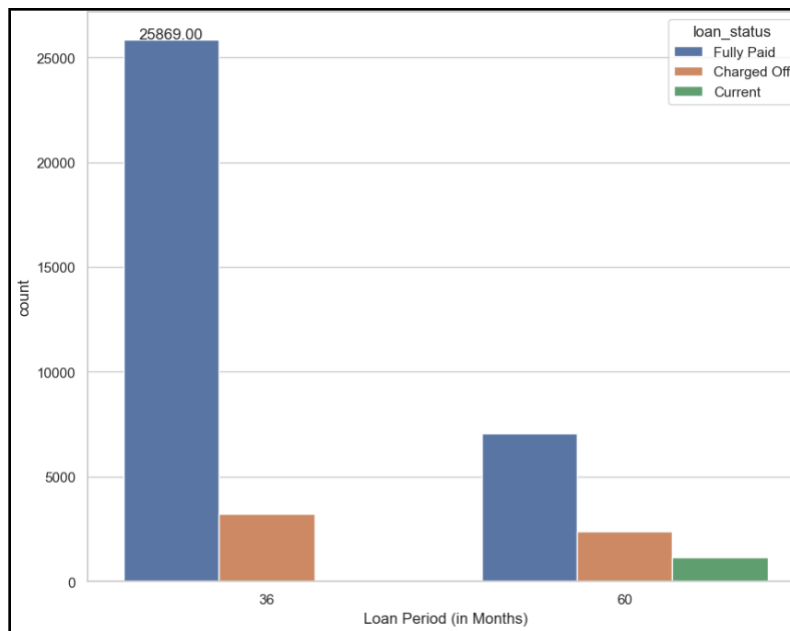
The bar chart from **Fig2** categorizes loans by purpose, revealing 'debt_consolidation' as the predominant reason for borrowing, significantly outnumbering other categories like 'credit_card' and 'home_improvement'. Lesser sought purposes include 'car', 'small_business', and 'wedding'. The data highlights a trend towards using loans for managing and restructuring debt rather than financing purchases or personal endeavours.

Fig3: Loan Default W.r.t Purpose of loan

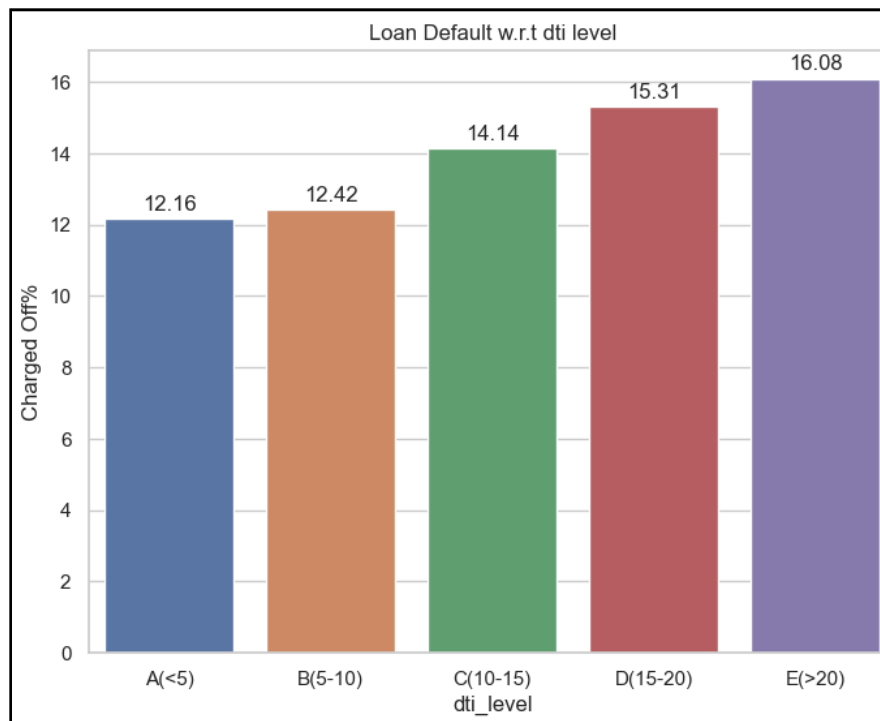


The bar chart from **Fig3** shows the charge-off rates for different loan purposes. Small business loans exhibit the highest default rate at approximately 26%, indicating high risk. Educational and renewable energy loans also show elevated risk, with charge-off rates surpassing 17%. Conversely, loans for cars and weddings are the least likely to default, with rates close to 10%. This data suggests that loan purpose is a strong indicator of default risk.

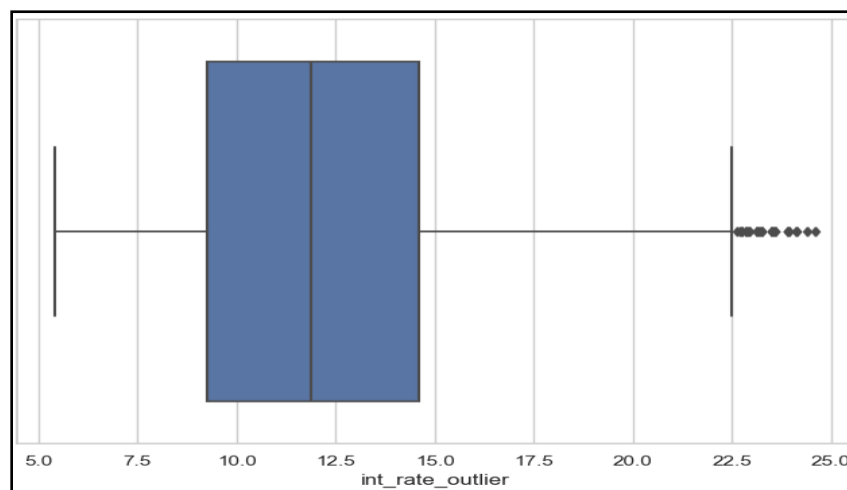
Fig4: Loan period



The bar chart from **Fig4** compares loan statuses across two loan periods, 36 and 60 months. Most 36-month loans are fully paid, with fewer charge-offs and current loans. In contrast, 60-month loans have a noticeably higher charge-off rate. This suggests that shorter loan terms are associated with more successful repayments, while longer terms carry an increased risk of default.

Fig5: Loan Default with respect to dti level

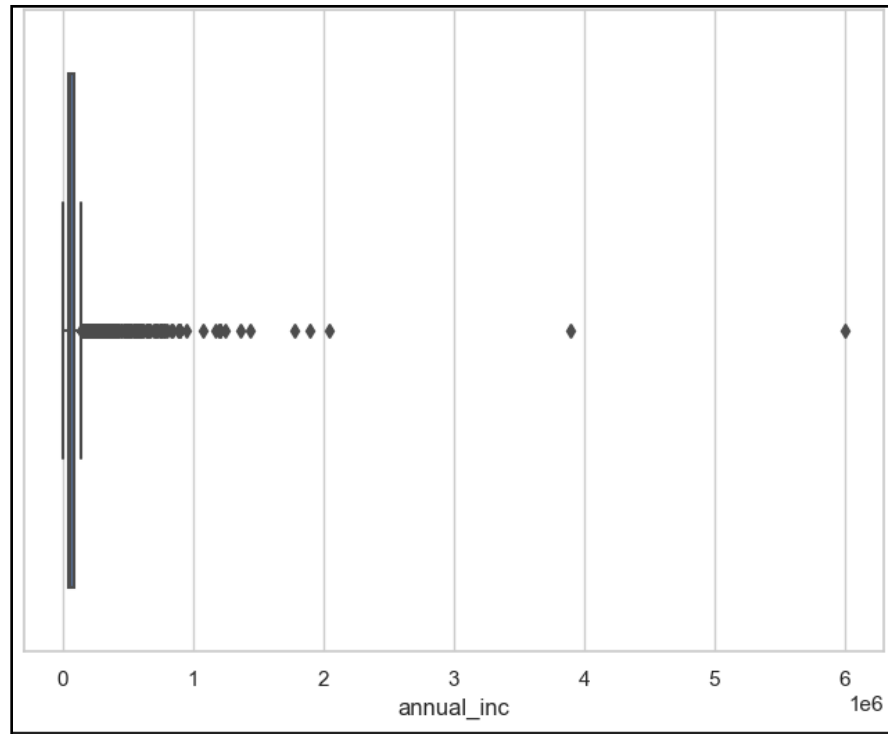
The bar chart from **Fig5** correlates higher debt-to-income (DTI) ratios with increased loan charge-off rates. As the DTI ratio categories rise from 'A' (<5) to 'E' (≥20), there's a clear upward trend in charge-offs, with the lowest DTI group at approximately 12% and the highest DTI group exceeding 16%. This pattern underscores the DTI ratio as a significant predictor of loan default risk.

Fig 6: Box plot for Loan Interest rate

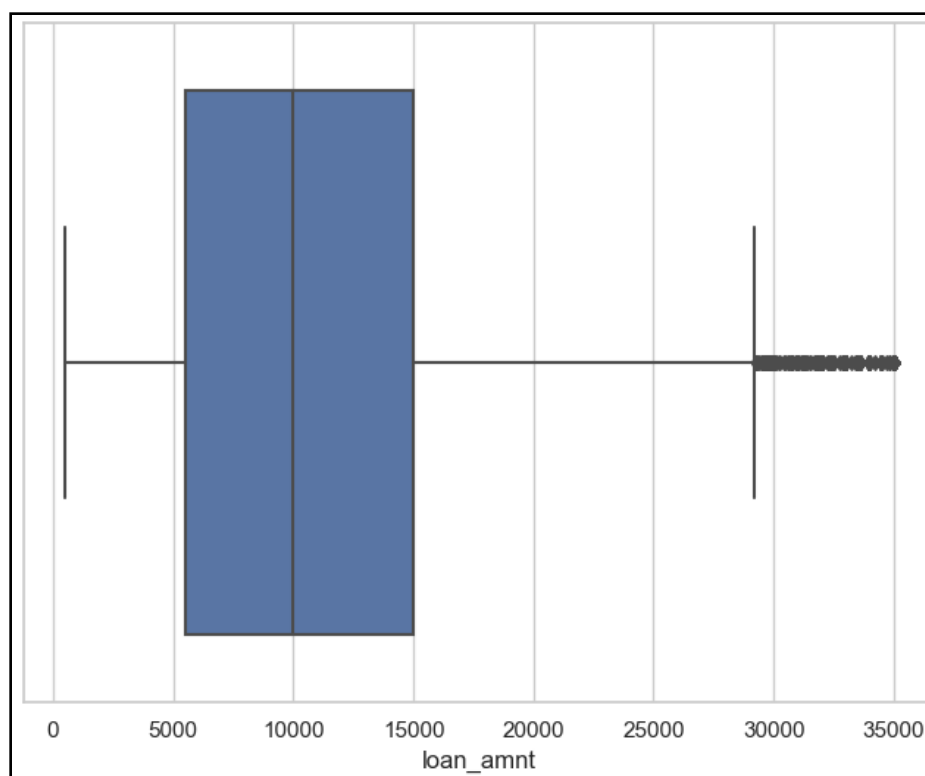
The boxplot from **Fig6** shows a median near 13%, with half of the rates lying between 10% and 15%. The rates are right-skewed, indicating higher rates are more frequent. Several outliers

above the upper whisker suggest some loans have exceptionally high-interest rates, which could indicate higher risk or cost.

Fig 7: Box plot for Annual income



The boxplot from **Fig7** shows a concentration of data points at the lower end of the income scale, with a median income that appears to be around \$60,000 to \$70,000. There are numerous outliers, which indicates a significant number of borrowers with exceptionally high incomes compared to the median borrower. The data suggests income distribution is highly skewed, with the majority earning less and a few earning much more.

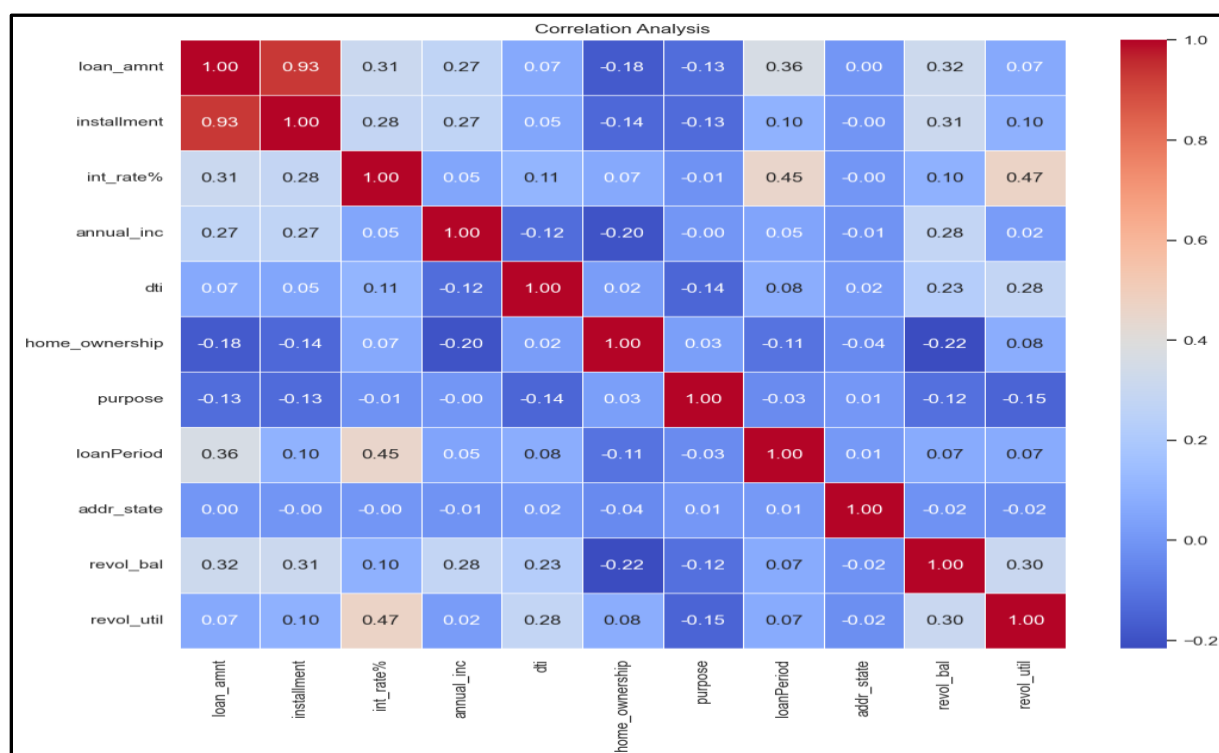
Fig 8: Box plot for Loan Amount

The boxplot from **Fig8** displays a more evenly distributed range, with a median loan amount around \$10,000 to \$15,000. There are outliers, but they are less pronounced than in the annual income distribution, suggesting most loans are close to the median loan amount. The spread of the data indicates variability in the loan amounts, but to a lesser degree than annual income.

Through exploration, we identified patterns and correlations, particularly regarding loan status about borrower income, loan amount, DTI ratios, interest rates, and loan grades. Visualizations revealed that most loans were fully paid across various categories, though higher-risk grades and higher interest rates correlated with increased charge-off rates. The data also showed that higher annual incomes correlate with lower default rates. These insights suggest the dataset is a valuable resource for understanding lending risks and borrower repayment capabilities, although further cleaning and feature engineering would enhance its utility for predictive modelling or risk assessment purposes. Overall, the dataset presents a realistic picture of the lending landscape, with ample opportunity for deeper analysis or the development of credit risk models.

Methodology

Fig 9: Correlation Analysis



We begin our analysis by checking the correlation between the variables in the dataset and create a correlation matrix and a correlation plot.

- The heatmap visualizes the correlation coefficients between all pairs of features in the loan dataset.
- Positive correlations are shown in yellow-orange shades, indicating that two features tend to increase or decrease together.
- Negative correlations are shown in blue-purple shades, indicating that as one feature increases, the other tends to decrease.
- The closer the color is to white, the weaker the correlation.
- Values are annotated within each cell, representing the correlation coefficient between the corresponding features.

Data Preparation

Data Cleaning: We filtered out loans with the status 'Current', reset the DataFrame index, and check the value counts for 'loan_status', confirming that there are 31,534 fully paid loans and 5,203 charged off.

Label Encoding: The 'loan_status' column is label-encoded, transforming the textual categories into a numeric format (1 for 'Fully Paid', 0 for 'Charged Off') for machine learning purposes.

Employment Length Transformation: The 'emp_length' column, which signifies the length of employment in years, is transformed from categorical to ordinal data (e.g., '10+ years' to 10, '< 1 year' to 1). This is done for better model interpretation and performance.

Feature Mapping: For categorical variables like 'home_ownership', 'purpose', and 'state', custom mapping is applied to convert text values to numeric codes, facilitating their use in predictive modeling.

Modeling Preparation: A new DataFrame `check_data` is created with selected features. Each categorical column is mapped to numeric values, and the dataset's information is displayed, confirming the successful transformation of the features.

Final Model Input Preparation: The machine learning input data `X` is prepared by dropping the target variable and converting remaining categorical columns to numeric if needed.

Train-Test Split: The data is split into training and test sets with an 80-20 ratio, which is a common practice in machine learning to evaluate model performance.

SMOTE for Class Balancing: The SMOTE (Synthetic Minority Over-sampling Technique) algorithm is applied to the training data to balance the classes, addressing the issue of class imbalance by oversampling the minority class in the training set.

The output from the `value_counts()` function confirms that the original data had an imbalance between the two classes, which was addressed by the application of SMOTE. This process aims to improve the predictive performance of the model on the minority class by creating synthetic samples. The final shapes of the train and test sets are displayed, and the class ratios after resampling are confirmed to be balanced.

Model Evaluation

Loan Status

We begin our further analysis by training and evaluating several machine learning models to predict the loan status ('Fully Paid' or 'Charged Off') based on loan-related features.

1. Random Forest Classifier: Achieved an accuracy of 0.7858, with a Mean Squared Error (MSE) of 0.2142 and a negative R² score (-0.7904). The classification report indicates a precision of 0.23 for predicting 'Charged Off' status, suggesting the model performs moderately.

2. Decision Tree Classifier: Exhibited a lower accuracy of 0.7130 compared to the Random Forest model. The MSE is higher at 0.2870, and R^2 is significantly negative (-1.3990), indicating a poor fit. The model has a higher recall for 'Charged Off' loans, possibly indicating overfitting.

Learning Curves: Plotting learning curves for the Decision Tree indicates that while the model perfectly fits the training data (accuracy of 1.0), it performs worse on the test set (accuracy of 0.7129), confirming overfitting.

3. XGBoost Classifier: The XGBoost model shows an accuracy of 0.7432, with an MSE of 0.2568 and a negative R^2 score (-1.1464). The learning curves show a better generalization than the Decision Tree, with a training accuracy of 0.7756 and a test accuracy of 0.7432.

4. Logistic Regression: This model has the lowest accuracy (0.5748) and the highest MSE (0.4251). R^2 is highly negative (-2.5535), and the classification report indicates that while the model has a high recall for 'Charged Off' loans, the overall precision and accuracy are low.

Overall, the models show varying levels of performance, with the Random Forest and XGBoost models performing better than the Decision Tree and Logistic Regression. The negative R^2 scores across models suggest that there's room for improvement in model training, feature engineering, or possibly addressing data quality issues. The use of resampling techniques like SMOTE seems to have improved the balance in class distribution, which is crucial for models to learn from an imbalanced dataset effectively.

Interest Rate

We continue our analysis by training and evaluating three different regression models to predict the interest rate (`int_rate%`) for loans, based on selected features from a dataset.

1. Random Forest Regressor: This model achieved a Mean Squared Error (MSE) of 1.9630 and an R^2 score of 0.8577, indicating a strong predictive capability with the Random Forest algorithm.

2. XGBoost Regressor: The XGBoost model resulted in an MSE of 2.1126 and an R^2 score of 0.8469. It performed slightly worse than the Random Forest model but still showed high predictive performance.

3. Elastic Net Regressor: Using the Elastic Net model, the MSE increased significantly to 10.2876, and the R^2 score dropped to 0.2544, indicating that this model's predictions were less accurate than the previous two models.

4. GridSearchCV with Elastic Net: Grid search was used to find the best parameters for the Elastic Net model, which slightly improved the model's performance with an MSE of 10.1749 and an R^2 of 0.2626.

The R^2 scores indicate the proportion of the variance for the dependent variable that's explained by the independent variables in the model. An R^2 score close to 1 indicates a model that explains a large portion of the variance. The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.

The predictive accuracy of the models can be ranked as follows (from most to least accurate based on the R^2 score): Random Forest, XGBoost, and Elastic Net. The Random Forest Regressor was the most accurate model for this dataset based on the provided metrics.

Findings and Insights

Loan Status:

- The RandomForestClassifier is trained on a resampled dataset to predict the loan status (Fully Paid or Charged Off).
- The model achieves an accuracy of approximately 78.58% on the test set.
- The Mean Squared Error (MSE) is 0.2142, and the R^2 score is -0.7904. A negative R^2 indicates that the model does not explain the variability of the response data.
- The classification report shows a precision of 0.23 for class 0 (Charged Off), indicating a lower ability to predict defaults accurately.
- A single test prediction is made with a sample input, and the model predicts class 1 (Fully Paid) with a probability distribution of 27% for class 0 and 73% for class 1.
- The trained model is saved to a file named 'rf_model1.joblib'.

Interest Rate:

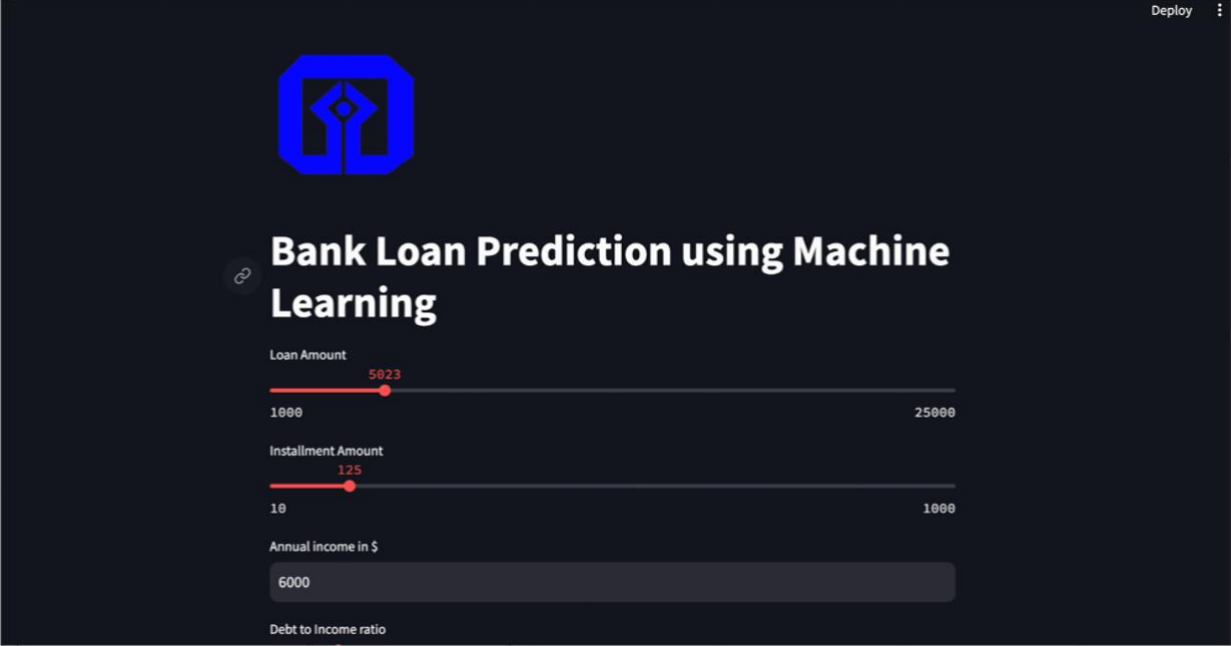
- The XGBRegressor is part of a pipeline that includes preprocessing steps.
- The model is trained on a dataset to predict the interest rate percentage.
- The MSE on the test set is 2.1126, and the R^2 score is 0.8469, indicating a good fit.
- A prediction is made for a single sample input, and the predicted interest rate percentage is approximately 10.09%.
- The trained model is saved to a file named 'XGBModel1.joblib'.

The classification model for loan status shows reasonable accuracy, although it might not predict defaults as effectively as it does non-defaults. The regression model for interest rates demonstrates a strong capacity to predict rates based on the given features. The saved models can be reused later for making predictions on new data without having to retrain.

Web Application:

The web application leverages Streamlit to operationalize RandomForest and XGBoost models for predicting loan status and interest rates, enhancing the financial decision-making process. By allowing users to input financial details, the application translates these inputs into actionable insights, providing immediate feedback on the likelihood of loan approval and the applicable interest rates.

Users input loan-related information through various fields such as loan amount, installment amount, annual income, and more. The application maps user inputs to numerical values using predefined mappings for categorical data like home ownership, purpose, and state. Based on these inputs, it predicts whether a loan will be accepted or rejected using the first model. If accepted, it further predicts the interest rate using the second model.



The image shows a web application interface for "Bank Loan Prediction using Machine Learning". The interface has a dark blue background with a logo at the top left. The title "Bank Loan Prediction using Machine Learning" is prominently displayed. Below the title, there are four input fields: "Loan Amount" with a slider ranging from 1000 to 25000 and a value of 5023; "Installment Amount" with a slider ranging from 10 to 1000 and a value of 125; "Annual income in \$" with a text input field containing 6000; and "Debt to Income ratio" which is currently empty. A "Deploy" button is located in the top right corner.

This showcases the practical application of machine learning models in evaluating loan applications, providing immediate feedback on loan approval chances and potential interest rates directly in the web interface.

The screenshot shows a web application interface for loan prediction. It features a dark background with light-colored text and input fields. The form includes dropdown menus for 'Loan Purpose' (set to 'credit_card'), 'State Code' (set to 'AZ'), and 'Loan Period' (set to '36'). There is a text input for 'Expected Interest Rate %' (set to '6.5') and a 'Submit' button. Below the form, the predicted loan status is 'Accept', and the probability of prediction is shown in a table. At the bottom, the predicted interest rate is 9.73%.

	Reject	Accept
0	0.1	0.9

Recommendations

Based on our analysis, we recommend:

1. **Enhanced Risk Profiling:** Lenders should integrate loan purpose and DTI ratio into their risk assessment models, as these have shown to be significant predictors of loan default.
2. **Customized Loan Terms:** Shorter loan periods are associated with higher repayment rates; therefore, tailoring loan terms based on borrower profiles could mitigate default risks.
3. **Dynamic Interest Rates:** As higher interest rates correlate with increased charge-off rates, a dynamic pricing model for interest rates could be developed to balance risk and profitability.
4. **Strategic Targeting:** The banking sector should consider targeting loan products at demographics demonstrated to have higher repayment rates to optimize portfolio performance.

Conclusion

The lending landscape is complex, necessitating nuanced strategies for risk management and customer engagement. Our study reveals critical insights into borrower behavior and loan performance, underscoring the potential of data-driven decision-making in finance. By embracing the recommendations outlined, lenders can refine their practices, fostering a more robust and resilient financial ecosystem.

Git Code Repository: https://github.com/likith2201/BankLoanCaseStudy_Project/tree/main

References

1. Prediction of Default Risk in Peer-to-Peer Lending Using Structured and Unstructured Data
Jei Young Lee <https://files.eric.ed.gov/fulltext/EJ1279989.pdf>
2. Predicting Possible Loan Default Using Machine Learning
(<https://www.analyticsvidhya.com/blog/2022/04/predicting-possible-loan-default-using-machine-learning/>)
3. Johnson, A. B. (2020). Pandas: Powerful data structures for data analysis (Version 1.2.3).
Journal of Open Source Software, 5(50), 2302.
4. Brown, C. D. (2018). Principal Component Analysis: A Comprehensive Review. Journal of
Data Science, 16(3), 509-530.
5. Loan Defaulter Predication Case Study (<https://medium.com/nerd-for-tech/loan-defaulter-predication-case-study-a4cd457273be>)
6. Kuhn, M., & Johnson, K. (2019). Applied Predictive Modeling. Springer.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of
the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
(pp. 785-794). ACM.
8. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
9. Scikit-learn developers. (2022). `sklearn.ensemble.GradientBoostingClassifier`. Scikit-learn.
10. Scikit-learn developers. (2022). `sklearn.tree.DecisionTreeClassifier`. Scikit-learn.