Date: 22.11.2024

Consulting Report for Module 5

Prepared by: Marina Chekulaeva

Contact: mchekula@gmail.com

For: Aref Hosseini

Project: Predicting Loan Repayment using LendingClub Data

---

1. Project Overview

Aref Hosseini's Module 3 project focuses on analyzing data from LendingClub, an online financial platform that connects borrowers with individual lenders. The primary objective of the project is to predict whether a loan will be fully paid back based on various features provided in the dataset.

The key steps undertaken in the project include:

- **Exploratory Data Analysis (EDA):** Investigating correlations between different features to understand their relationships.
- **Visualization:** Using scatter plots and histograms to visualize the distributions and correlations of features like interest rate (int.rate), FICO score (fico), and others.
- **Model Implementation:** Implementing Logistic Regression, Decision Tree Classifier, and Random Forest Classifier to build predictive models.
- **Evaluation:** Assessing model performance based on precision and accuracy metrics.

---

2. Positive Aspects and Strengths

Exploratory Data Analysis

- **Identification of Data Imbalance:** Through EDA, the imbalance in the dataset was identified—443 fully paid loans versus 2,431 not fully paid loans. Recognizing this imbalance is crucial as it significantly impacts model performance.
- **Correlation Analysis:** Starting with Pearson correlation plots to examine pair-wise correlations among features is an effective approach to identify relationships. The observed anti-correlation between interest rate and FICO score provides valuable insights into how credit scores affect interest rates.
- **Visualization Techniques:** Utilizing scatter plots and histograms to compare the distributions of FICO scores between fully paid and not fully paid loans helps in visually assessing whether certain features influence loan repayment.

Model Implementation

- **Choice of Algorithms:** Selecting Logistic Regression for initial modeling is appropriate for binary classification problems. The use of Decision Tree Classifier

and Random Forest Classifier further allows capturing complex, non-linear relationships between features.

- **Model Performance Evaluation:**
  - **Confusion Matrix Analysis:** Aref Hosseini employed confusion matrices to evaluate model performance. The confusion matrices revealed high false negatives, indicating that many payers were predicted as non-payers.
  - **Critical Observations:** Recognizing that the models predict most of the non-payers correctly but struggle with accurately predicting payers demonstrates an understanding of the impact of data imbalance on model performance.

---

3. Suggestions for Improvement

Enhanced Exploratory Data Analysis

- **Split Data by Class for Analysis:** It would be beneficial to split the dataset into fully paid and not fully paid loans during EDA. This allows for a detailed comparison of feature distributions between the two classes.
  - **Feature Distribution Analysis:** Examine histograms, box plots, and statistical summaries (mean, median, standard deviation) of each feature for both classes to identify which features differ significantly and may contribute to loan repayment prediction.
  - **Statistical Testing:** Perform statistical tests (e.g., t-tests for continuous variables, chi-square tests for categorical variables) to determine if differences in feature distributions between the classes are statistically significant.

Addressing Data Imbalance

- **Resampling Techniques:** To mitigate the effects of class imbalance, consider applying resampling methods:
  - **Oversampling:** Increase the number of instances in the minority class (fully paid loans) using techniques like SMOTE (Synthetic Minority Over-sampling Technique).
  - **Undersampling:** Reduce the number of instances in the majority class (not fully paid loans) to balance the dataset.
- **Use of Specialized Algorithms:** Explore algorithms designed to handle imbalanced data, such as:
  - **Balanced Random Forest:** Adjusts the class distribution in each bootstrap sample.
  - **Cost-Sensitive Learning:** Assigns a higher cost to misclassifying the minority class, prompting the model to pay more attention to it.

4. Conclusion

Aref Hosseini has demonstrated a solid understanding of predictive modeling and data analysis techniques. By addressing the class imbalance and incorporating more robust statistical analyses and evaluation metrics, the predictive accuracy of the models can be significantly improved. Focusing on feature engineering and proper model validation will enhance the reliability of the conclusions drawn from the data.