Introduction: Overview of the data, predictive task, and summary findings

I was scrolling through datasets on Kaggle and came across a "loan dataset" sourced from LendingClub, a popular online lending platform that connects borrowers seeking loans with investors looking for profitable returns. For my final project, I aim to develop a classification model that can accurately predict whether a borrower will fully repay their loan or default, based on a variety of borrower financial profiles and loan characteristics—such as FICO score, debt-to-income ratio, credit history, and others, in relation to the target variable, 'not.fully.paid,' which indicates loan repayment behavior (with '1' for not repaid and '0' for fully repaid). This model will help financial institutions, like LendingClub, reduce the risk of loan defaults, improve the app's efficacy and reliability, and enable more informed decisions for lenders and investors.

Summary Findings:

Overall, the models demonstrated strong performance in predicting fully repaid loans, indicating that the features utilized in the dataset effectively captured patterns associated with reliable borrowers. However, a consistent issue emerged across all models: they struggled to accurately identify loans at risk of default. This limitation suggested that the dataset contained fewer prominent characteristics that distinguished defaulting borrowers from those who repaid their loans, making it difficult for the models to detect these high-risk cases. To address this challenge, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to balance the dataset significantly improving the models' ability to identify defaults. For example, in the XGBoost model, the recall for defaults increased from 8% to 26%, while the recall for defaults in the LightGBM model improved from 1% to 18%. These improvements enabled the models to

capture a greater proportion of high-risk loans, thereby enhancing their effectiveness in identifying potential defaults. Since all of these models were successful in identifying those who repaid the loans, I decided to evaluate the models based on Recall and F-1 Score, as both provide insights specifically on defaults. At the end, the logistic regression model emerged as the most effective one with an F-1 score of 0.26 and a recall of 51.8% for the default class. While the precision for predicting defaults remained low at 17.7%, the model's ability to correctly identify more than half of the actual defaults demonstrated its capacity to recognize high-risk borrowers. This combination of metrics indicates that logistic regression offers a reliable and interpretable approach to predicting borrower outcome, making it a valuable tool for lending institutions, such as LendingClub, capturing as many default cases as possible and thus minimizing financial loss/risk.

Data Description: Data source and description

The dataset for this project was sourced from Kaggle in CSV format and provides detailed information about loans issued through LendingClub. It includes borrower financial profiles, loan characteristics, and repayment outcomes. Notably, the dataset contains no missing values, allowing us to focus on feature engineering and exploratory analysis without the need for extensive data cleaning. The dataset includes the following key features:

- **credit.policy**: A binary indicator (1 or 0) showing whether the borrower meets LendingClub's credit criteria.

- **purpose**: A categorical variable representing the intended use of the loan (e.g., credit card, debt consolidation, education).

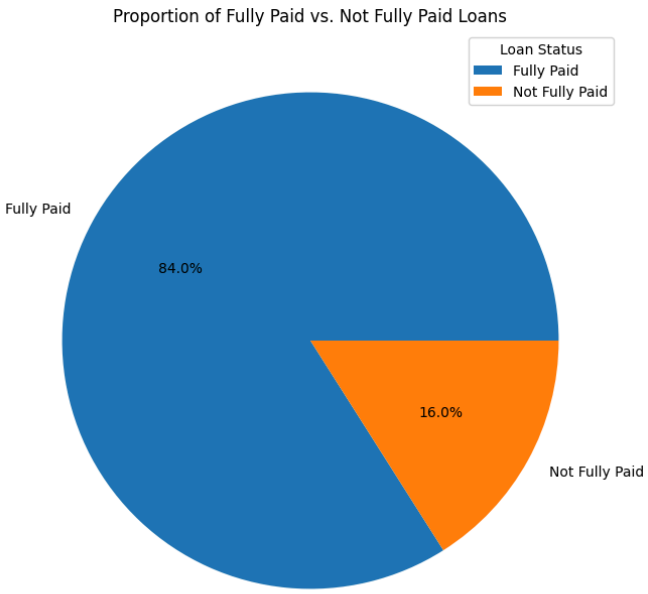
- **int.rate**: The interest rate of the loan, with higher rates generally assigned to riskier borrowers.
- **installment**: The monthly repayment amount for the borrower.
- **log.annual.inc**: The natural logarithm of the borrower's reported annual income.
- **dti**: The debt-to-income ratio, indicating the borrower's financial stability.
- **fico**: The borrower's FICO credit score, a measure of their creditworthiness.
- **days.with.cr.line**: The number of days the borrower has had a credit line.
- **revol.bal**: The outstanding balance on the borrower's revolving credit accounts.
- **revol.util**: The percentage of the borrower's available credit that is being utilized.
- **inq.last.6mths**: The number of credit inquiries made by creditors in the past six months.
- **delinq.2yrs**: The number of times the borrower was 30 or more days late on a payment in the past two years.
- **pub.rec**: The number of derogatory public records, such as bankruptcies or tax liens.

The main target variable in this dataset is whether or not a borrower has fully repaid their loan, which is labeled as **not.fully.paid**. This variable indicates if the borrower has completely repaid the loan (0) or defaulted in part or full (1). As we proceed with the analysis, we will focus on feature selection, model training, and evaluation.

Models and Methods: Overview of models and implementation

First, I began with **Exploratory Data Analysis (EDA)** to understand the relationships between the independent variables (*X*) and the dependent variable (*Y*), which indicates whether the loan was fully paid or not. To gain a general overview, I created a **pie chart** showing the proportion of fully paid versus non-fully paid loans. The chart revealed that **84%** of loans were fully paid, while **16%** were not (i).

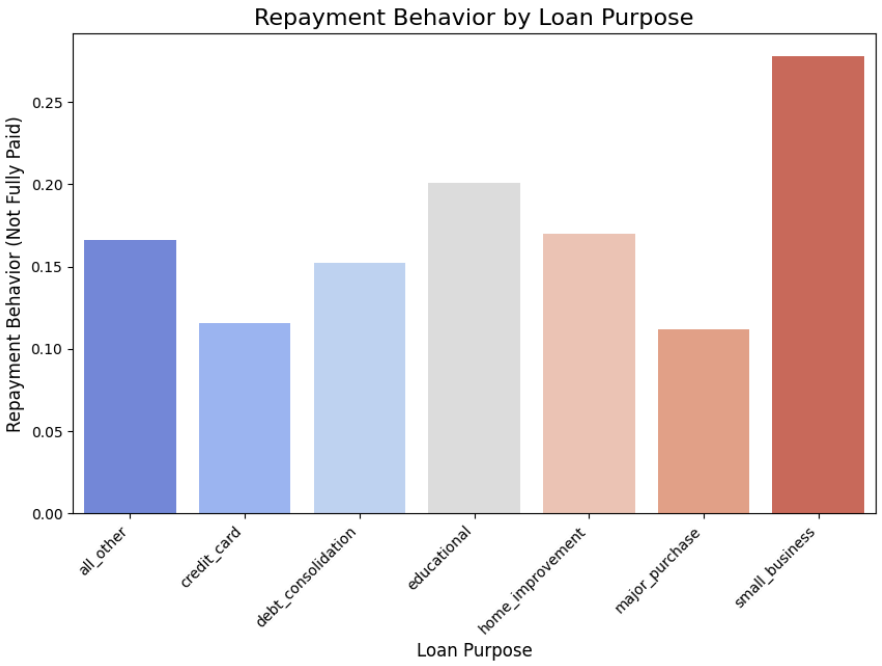
(i)



Next, I analyzed the only categorical variable in the dataset, **'purpose'**, which represents the reason for taking out the loan. Seven unique loan purposes were found, namely: 'debt_consolidation,' 'all_other,' 'credit_card,' 'home_improvement,' 'small_business,' 'major_purchase,' and 'educational.' To visualize the relationship between loan purpose and repayment behavior, I generated a bar plot grouped by 'purpose' to display the average repayment behavior (ii). This analysis showed significant differences based on loan purposes.

For example, loans for **'small_business'** exhibited a higher likelihood of non-repayment, while loans for **'major_purchase'** had a lower likelihood of non-repayment. Therefore, 'purpose' is an important independent variable for predicting the dependent variable **'not.fully.paid'**, and should be taken into account in our analysis.

(ii)

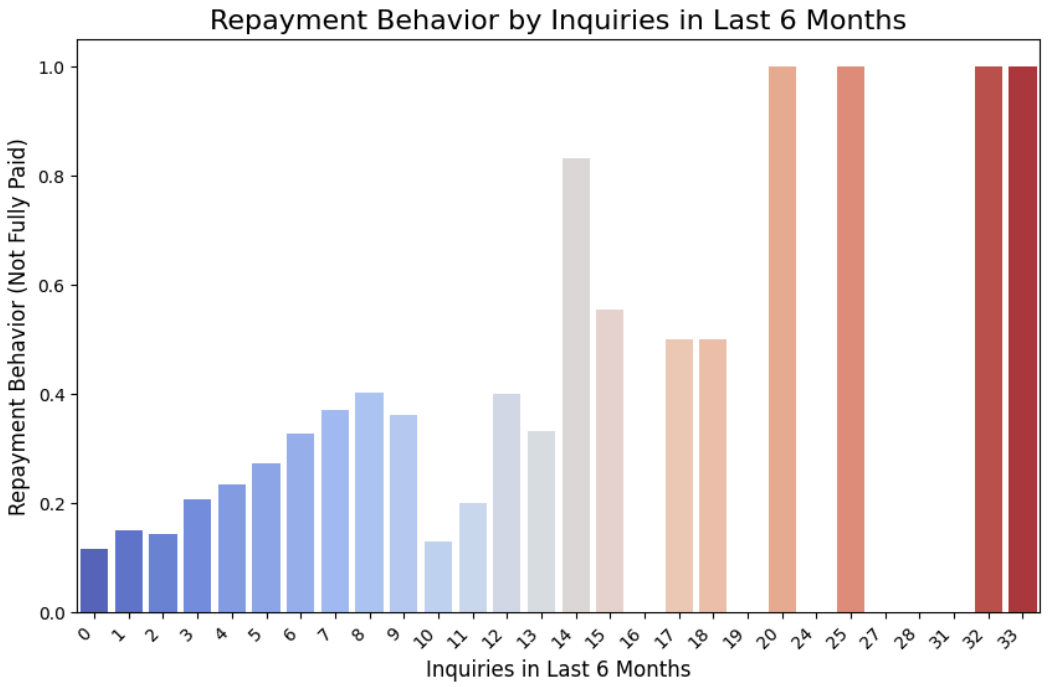


Continuing the analysis, I examined the relationships between other numeric independent variables and repayment behavior ('not.fully.paid'):

- **'inq.last.6mths'**: This variable indicates the number of credit inquiries in the past six months. I grouped by 'inq.last.6mths' and calculated the average repayment behavior. A high number of inquiries often suggests financial instability and a higher risk of

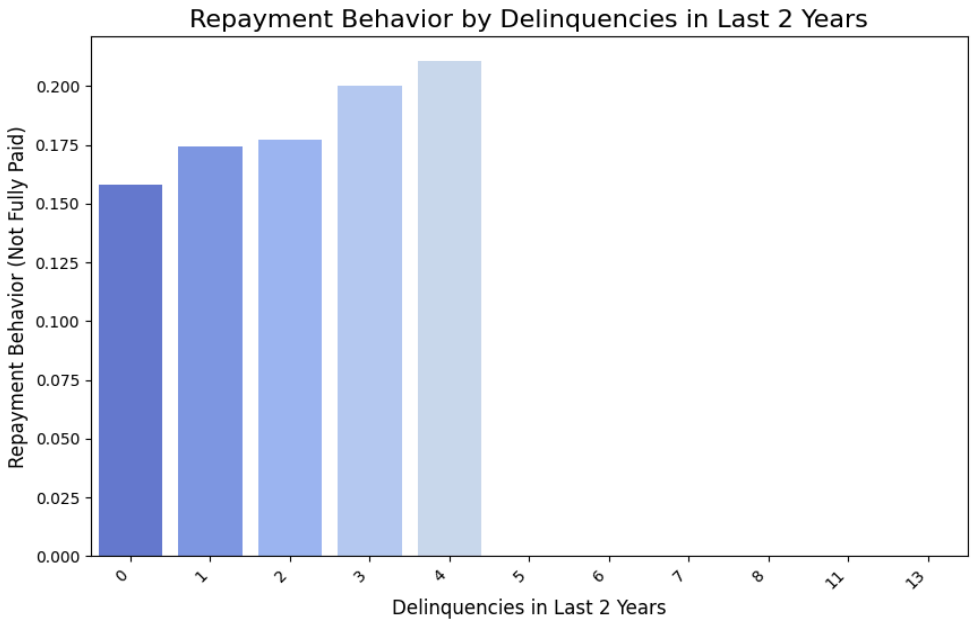
non-repayment (iii). However, further analysis revealed that the correlation was weaker than expected, possibly due to data sparsity or inconsistencies in the variable distribution.

(iii)



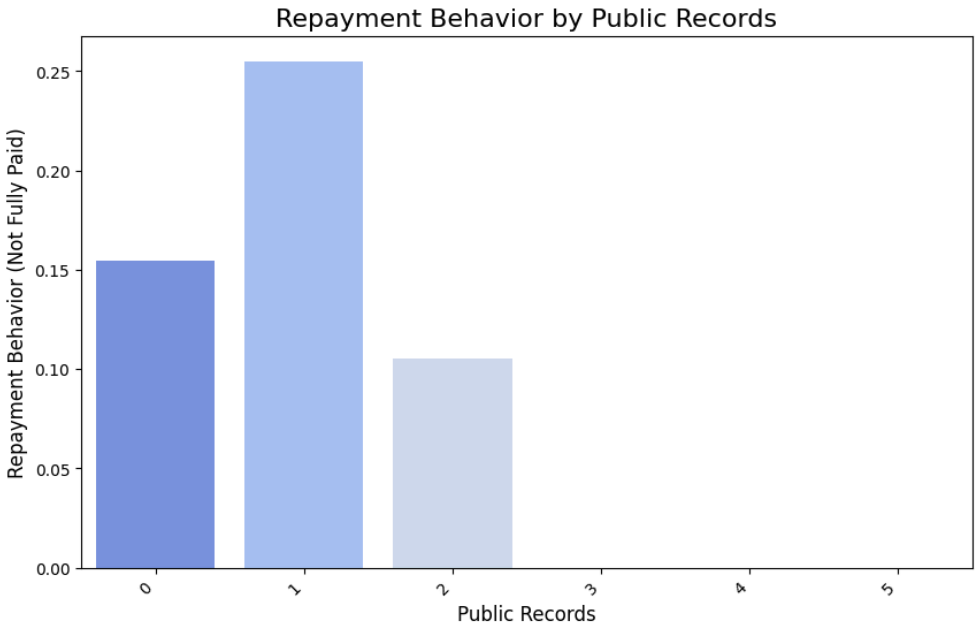
- **'delinq.2yrs':** This variable represents the number of times the borrower was more than 30 days late on a payment in the past two years. Grouping by 'delinq.2yrs' and calculating average repayment behavior helped identify whether frequent delinquencies correlated with non-repayment. It was evident that a higher number of delinquencies was associated with an increased likelihood of non-repayment, supporting the intuition that borrowers with a history of missed payments are more likely to default (iv). However, this pattern began to break down once the number of delinquencies reached 5-13.

(iv)



- **'pub.rec'**: This variable counts derogatory public records, such as bankruptcies or tax liens. I grouped by 'pub.rec' to observe its relationship with loan repayment, expecting that more public records would correlate with a higher likelihood of non-repayment. Interestingly, the barplot showcased an unusual pattern where borrowers with 3-5 public records were more likely to fully repay their loan, while those with 0-2 records were more likely to default (v). This anomaly may be due to data limitations or underreporting in certain categories, which requires further investigation.

(v)



Given these insights, I decided to drop all three variables from the analysis as they did not contribute significantly to the model's predictive power or showed inconsistent patterns that could skew the results.

Finally, I created a **pair plot** to visualize relationships between the numeric variables **'fico'** (credit score), **'log.annual.inc'** (logarithm of annual income), **'dti'** (debt-to-income ratio), and the dependent variable **'not.fully.paid'**. The pair plot visualizes the relationships between the variables **fico**, **log.annual.inc**, **dti**, and **not.fully.paid**. From the plot, several key observations can be made (vi):

1. **fico vs. log.annual.inc**:
    - There doesn't appear to be a strong linear relationship between the credit score (**fico**) and the logarithm of annual income (**log.annual.inc**), as the points are spread across the plot.

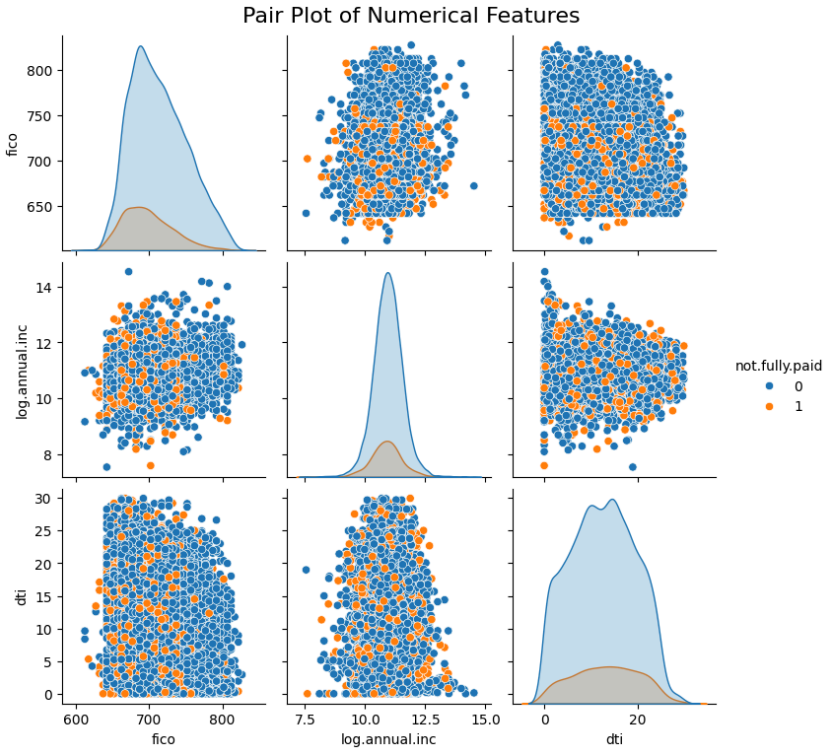
- The distribution for both **fico** and **log.annual.inc** shows a skewed pattern, with higher concentrations in the lower values for credit score and income.

2. **fico vs. dti**:
    - There is no clear pattern or correlation between **fico** and **dti**. Borrowers with high and low credit scores (**fico**) seem to have a range of **dti** values.
    - **dti** appears to be more spread out, with both low and high debt-to-income ratios present across credit scores.
3. **log.annual.inc vs. dti**:
    - A weak inverse relationship can be observed here, where lower annual incomes tend to coincide with higher debt-to-income ratios. This suggests that borrowers with lower incomes may be over-leveraged, which could be an important factor in loan repayment.
4. **not.fully.paid (color coding)**:
    - The **not.fully.paid** variable (depicted in orange for non-repayment) is distributed across all values of **fico**, **log.annual.inc**, and **dti**, which suggests that loan repayment behavior isn't strictly tied to these financial factors alone.
    - There does not seem to be a clear separation between fully paid (blue) and not fully paid (orange) loans based on the **fico** and **log.annual.inc** variables, although there is some indication that borrowers with lower credit scores and higher debt-to-income ratios may be more likely to not fully repay their loans.

Overall, the pair plot offers valuable insights into the relationships between the financial features and repayment behavior. While no strong linear correlations are observed, it does suggest that

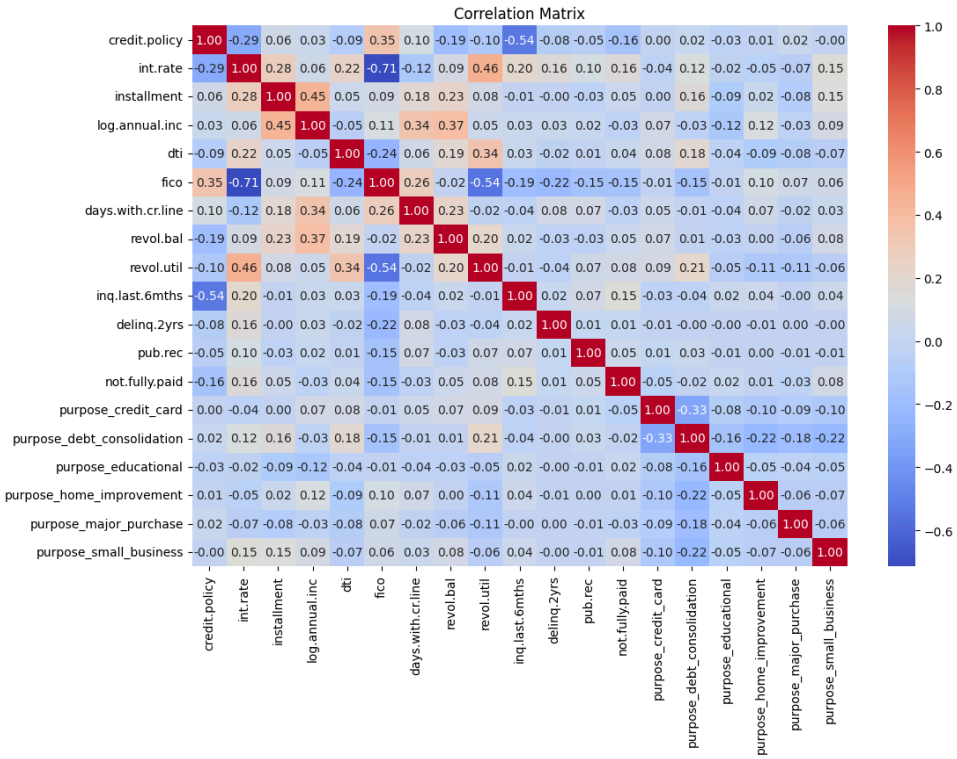
certain factors, like **dti** and **fico**, may have a more subtle influence on whether a loan is repaid in full. These patterns can guide feature selection and further analysis when building predictive models.

(vi)



Next, I created and analyzed a correlation matrix and found some notable relationships. For example, **fico** (credit score) has a negative correlation with **int.rate** (interest rate), suggesting that better credit scores are associated with lower interest rates. Additionally, **not.fully.paid** is weakly correlated with **fico**, indicating that lower credit scores may be linked to a higher likelihood of default. This matrix helps me identify important predictors and potential multicollinearity issues for further feature selection in the model.

(vii)



After my EDA, I began with feature engineering for my models. I first standardized the continuous features ‘dti’ (debt-to-income ratio) and ‘installment’ (monthly repayment amount) using **StandardScaler** to ensure that these features are on the same scale, improving the performance of certain models. Next, I converted the ‘credit.policy’ and ‘not.fully.paid’ columns to boolean data types. This is because these columns represent binary values (1 or 0), and converting them to boolean values simplifies the modeling process. These transformations help ensure that the dataset is ready for modeling, with continuous variables standardized and categorical variables appropriately encoded.

Later, I defined the feature set XX and the target variable yy. The feature set XX includes all the columns except for the target variable, ‘not.fully.paid.’ The target variable yy is therefore ‘not.fully.paid.’ I then split the dataset into training and testing sets using ‘train_test_split’ from

sklearn, ensuring that the model would be trained on one subset of the data and tested on another, providing an accurate performance evaluation. I started by initializing a LogisticRegression model and fitting it to the training data (X_train, y_train). This allows the model to learn the relationship between the features and the target variable ('not.fully.paid'). Next, I used the fitted model to make predictions on the test data (X_test). The predicted values were stored in predict_y_lr. Same thing was performed for Random Forest (predict_y_rf), XGBoost (predict_y_xgb), and LightGBM (predict_y_lgb).

Results and Interpretation: Review of Modeling Results and Interpretation of Performance

To improve loan repayment predictions using the LendingClub dataset, I tested five models—Logistic Regression, Random Forest, KNN, XGBoost, and LightGBM—each with unique strengths and approaches in tackling my final project's objective. Logistic Regression is simple and easy to interpret, making it a good starting point/baseline model. Random Forest is great at picking up complex, non-linear relationships in the data while reducing the risk of overfitting. KNN helps by finding patterns based on similar borrowers, while XGBoost is known for its accuracy, especially when the data is imbalanced. LightGBM stands out for its speed and efficiency, making it perfect for handling large datasets with both continuous and categorical features. Together, these models provide a well-rounded approach to predicting loan repayment more accurately. The results of my models are as follows:

Logistic Regression Classifier:

The Logistic Regression model, although demonstrating high precision (84%) and perfect recall (100%) for predicting successfully repaid loans, showed a significant deficiency in detecting

defaults, with a recall of 0% for this class. This stark disparity indicates that while the model is effective in identifying loans likely to be repaid, it fails entirely to detect high-risk borrowers who are likely to default. This imbalance in performance suggests a crucial area for improvement, particularly in enhancing the mode''s sensitivity to defaults, potentially through methods like SMOTE to address the class imbalance issue.

Random Forest Classifier:

The Random Forest model showcased a high precision (84%) and recall (99%) in predicting loans that were fully repaid. However, it struggled considerably with loans that were not fully repaid, evidenced by a minimal recall (2%). Despite an overall accuracy of 83.8%, the model's low efficacy in identifying defaults underlines a critical gap, highlighting the need of addressing the data imbalance similar to the one seen in the Logistic Regression above.

KNN Classifier:

The KNN model showcased a high precision (84%) and recall (97%) in predicting loans that were fully repaid. However, it struggled considerably with loans that were not fully repaid, evidenced by a minimal recall (5%). Despite an overall accuracy of 82%, the model's low efficacy in identifying defaults underlines a critical gap, highlighting the need of addressing the data imbalance similar to the one seen in other models.

XGBoost Classifier:

The XGBoost model exhibited strong performance metrics for fully repaid loans with a precision of 85% and a recall of 96%. However, its ability to predict defaults was notably weaker, with a precision of 28% and a recall of only 8%. Despite an overall accuracy of 82.15%, the practical

utility of this model is limited by its poor detection of potential defaults, again, suggesting a need for adjustments in the model to enhance its detection of defaults.

LightGBM Classifier:

Similarly, the LightGBM model performed well in predicting loans that were fully repaid, achieving a precision of 84% and a recall of 99%. However, it showed significant shortcomings in accurately identifying loans that were not fully repaid, with the precision dropping to 16% and the recall to a low 1%, indicating a similar issue as observed in our previous models. The overall accuracy stood at approximately 83%, pointing to a strong ability to predict the majority class (fully repaid loans) but a notable deficiency in detecting loan defaults.

Overall, the analysis across these models indicates a consistent pattern: while all models are effective at identifying fully repaid loans, they falter significantly when it comes to detecting defaults. To effectively address class imbalance and potentially improve the detection accuracy of defaults, implementing SMOTE could help in refining the predictive capabilities of these models, thereby supporting more informed decision-making in lending scenarios.

Conclusion and Next Steps: Summary of models and next steps for further analysis

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.54	0.177	0.518	0.264
Random Forest	0.77	0.261	0.232	0.246
XGBoost	0.79	0.288	0.196	0.233
LightGBM	0.80	0.298	0.17	0.21
KNN	0.59	0.177	0.429	0.25

After implementing SMOTE to address the class imbalance, we observed a notable improvement in the models' ability to detect defaults:

- **Logistic Regression**: Before SMOTE, Logistic Regression struggled with detecting defaults, showing a recall of 0% for the minority class (defaults). After implementing SMOTE, the recall improved significantly to **51.8%**, indicating a better ability to identify high-risk borrowers. However, the precision remained low at **17.7%**, suggesting that many of the predicted defaults are false positives.
- **Random Forest**: The Random Forest model showed a notable improvement in detecting defaults after SMOTE, with recall rising to **23.3%**. The overall accuracy is **77.3%**, but the precision (26%) and recall for defaults remain relatively low, indicating the model still struggles to balance between detecting defaults and minimizing false positives.
- **KNN**: The KNN model showed moderate improvements in detecting defaults after SMOTE, with recall reaching **42.9%** and accuracy increasing to **59.2%**. However, the precision remained low at **17.8%**, indicating that the model is still prone to false positives.
- **XGBoost**: The XGBoost model performed well on predicting fully repaid loans, with precision and recall for this class reaching **28.8%** and **19.7%**, respectively. However, its ability to predict defaults remains weak, with precision at **28%** and recall at **8%**. The overall accuracy increased to **79.5%**, but the model still struggles with detecting defaults.
- **LightGBM**: The LightGBM model showed the highest accuracy (**80.4%**), but it still faced difficulties in detecting defaults, with precision dropping to **16%** and recall to **1%**

for the minority class. The model performed well for predicting fully repaid loans, but the detection of defaults remains a challenge.

I decided to evaluate the models using the F1-Score because it provides a balance between precision and recall, which is especially important for imbalanced datasets like ours, which still remains imbalanced even after implicating SMOTE. The F1-Score highlights how well the model can identify defaults, which is key for detecting high-risk borrowers. Based on this, Logistic Regression stands out as the best model for predicting loan defaults and high-risk borrowers.

Next Steps/Improvements:

To gain deeper insights into loan repayment behavior, I propose the following steps:

1. The analysis across these models consistently shows a similar pattern: the models excel in predicting fully repaid loans but face challenges in identifying defaults. Although SMOTE improved the detection of defaults, the models still struggle with high false-positive rates. To further improve the performance, additional techniques such as model tuning, ensemble methods, or alternative oversampling methods could be explored. Addressing this imbalance remains crucial for effectively predicting defaults and supporting more informed decision-making in lending scenarios.
2. Analyzing FICO Score Variables: First, I would investigate which variables currently contribute to the FICO score calculation. Understanding how each of these variables is weighted in the FICO score formula will provide a clear picture of their relevance in predicting loan repayment. For instance, knowing what percentage of the score comes

from credit history, amounts owed, and other factors would help identify the most influential components of the score for our model. After doing some research beyond this dataset, I found out that some of the independent variables in this dataset are already included when counting the FICO score. Thus, instead of counting the independent variables that are already a part of the FICO score, I would much rather prefer to compare other non-FICO related variables to the FICO score, identifying which variable is the best at predicting whether or not the borrower is going to repay the loan. This would be another interesting project to work on.

3. Investigating the Impact of Loan Purpose: The dataset contains a categorical variable indicating the purpose for which the loan was taken. Since the likelihood of repayment varies based on one's "purpose" for taking a loan, it would be interesting to do an in-depth analysis of the reasons or factors for which a certain 'purpose' has a higher/lower repayment rate than the other. Analyzing repayment trends across loan purposes could have practical implications for the LendingClub platform, as they could consider adjusting lending policies or 'credit policy' based on these insights. For example, the platform could implement stricter conditions for loans associated with 'small_business,' as these loans may carry a higher risk of non-repayment. By doing so, LendingClub could help both lenders and borrowers maintain accountability, reducing the likelihood of default and improving the overall efficacy/reliability of the platform.