Strategic Customer-Lender Matching: Analysis of Bankrate’s Loan Approval Data

1. Executive Summary

The Loan Prediction Analysis Project investigates the potential for optimizing Bankrate's personal loan customer-lender matching process. By analyzing customer data, we aim to identify key factors impacting loan approvals for different lenders (A, B, and C). Combining this with lender payout structures, we can develop a more targeted matching strategy that maximizes both loan approval rates and revenue generation for Bankrate.

1.1. Business Problem Statement

Currently, Bankrate matches customers with lenders based on general profiles and preferences. This approach may not be optimal for maximizing approvals and revenue. This project leverages customer data to identify approval drivers for each lender. Utilizing these insights, we can create a data-driven matching strategy that achieves both key business goals: high approval rates and increased revenue.

2. Methodology

2.1. Understanding the data structure

To understand more about the data structure, the dataset used is for analyzing personal loan applications from Bankrate.com, a consumer financial service company, subsidiary of Red Ventures. The dataset comprises 100,000 records and 14 columns, providing a substantial sample size for robust statistical analysis. This 100k records is the customer data to understand which factors matter the most in the loan approvability. Here's a detailed breakdown of the data structure:

2.1.1. Data Types:

We analyzed a substantial data set containing 100,000 records and 14 customer and loan details from Bankrate.com. This data allows for robust statistical analysis to understand factors influencing loan approval. The dataset incorporates various data types, including:

String: User ID, Reason, FICO Score Group, Employment Status, Employment Sector, Loaner

Integer: Application (always 1 per row), Loan Amount, FICO Score, Monthly Gross Income, Monthly Housing Payment, Bounty

Boolean: Ever Bankrupt or Foreclose, Approved

3. Descriptive Analysis

3.1. Unveiling Customer Trends

To optimize loan matching and gain insights into customer behavior, we begin with descriptive analysis. This initial phase involves exploring the data to uncover patterns, trends, and relationships between variables. It lays the groundwork for further analysis by providing a foundational understanding of the data before delving into more complex modeling techniques.

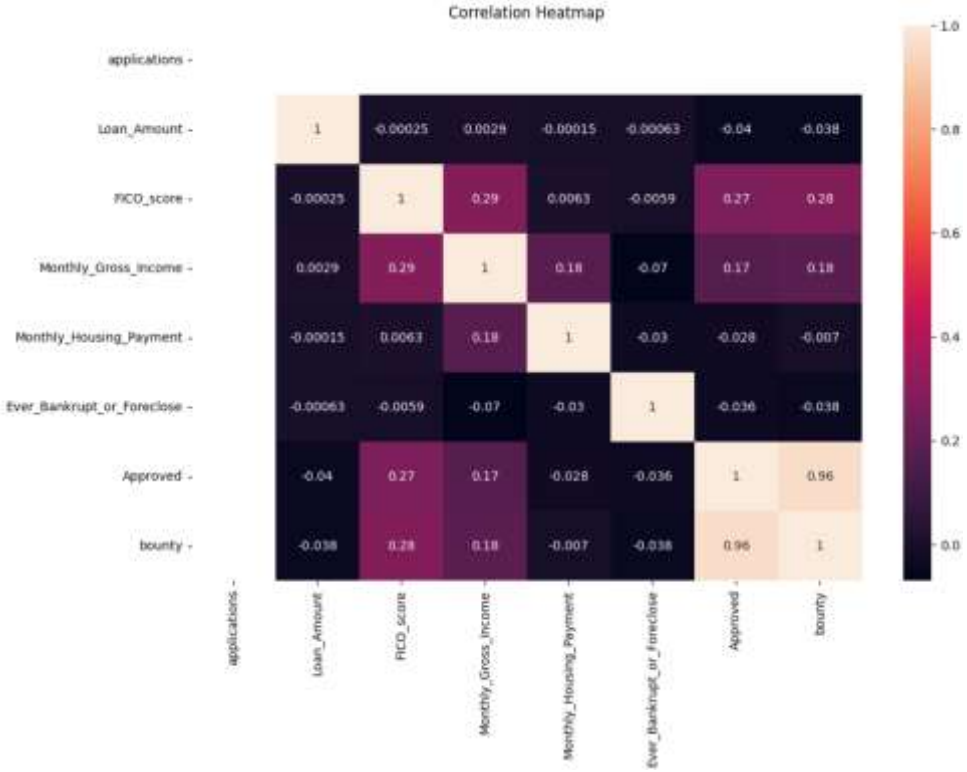
3.2. Data Quality Assessment

We employed the *data.info()* function to gain an overview of the data structure, including column names and data types. Additionally, we assessed the number of records in each column to identify potential inconsistencies. To check for missing values, we utilized *df.isnull().sum()*. The analysis revealed a clean data set with missing values only in the "Employment Sector" column. Since the number of missing values is minimal, we opted to treat these as null values for the analysis.

3.3. Data Exploration

3.3.1. Identifying Multicollinearity with Correlation Analysis

To understand the relationships between variables and identify potential multicollinearity, we employed correlation analysis. This involved calculating a correlation matrix, which quantifies the linear association between each pair of variables. We then visualized the correlation matrix using a heatmap. This visual representation allowed us to efficiently identify highly correlated variables that might contribute to multicollinearity in our models.



4. Data Analysis & Insights

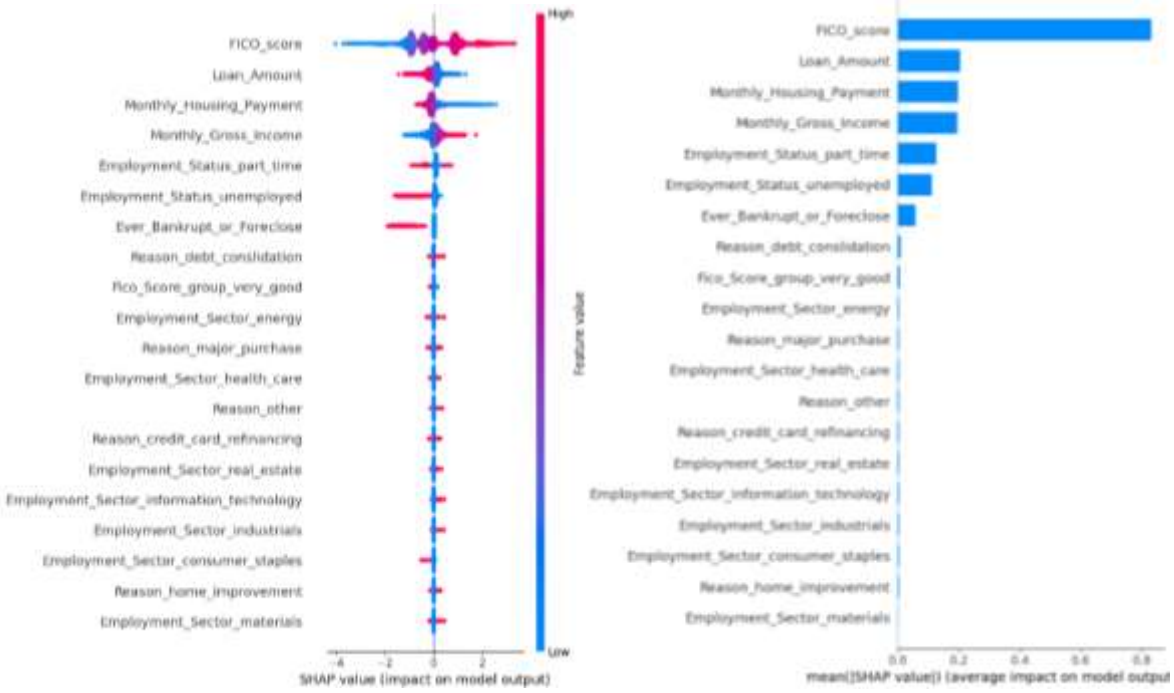
To address the core business challenge and identify the variables most influential in loan approval, we delved into the data through feature importance analysis.

4.1. Feature Importance with LightGBM

We employed Light Gradient Boosting Machine (LightGBM), a powerful machine learning algorithm renowned for its speed and accuracy. LightGBM excels at feature importance analysis, enabling us to identify the variables that most significantly impact loan approval decisions. By analyzing the importance scores assigned by LightGBM, we can determine which aspects of a customer's profile hold the greatest weight in the model's predictions.

4.2. SHAP Values: Transparency and Interpretability

To further enhance the interpretability and fairness of our analysis, we leveraged SHAP (SHapley Additive exPlanations) values. SHAP values provide a method for explaining individual predictions made by the LightGBM model. By examining these values, we can pinpoint the specific features within a customer's profile that have the most significant influence on the model's prediction of loan approval or denial. This granular level of insight allows us to understand not only which variables are important but also how they contribute to the final decision.



Insights from SHAP Analysis: Unveiling Key Predictors of Loan Approvability

By analyzing SHAP values, we gained valuable insights into the variables that most significantly influence loan approval decisions. Here's a breakdown of the key findings:

Strongest Predictors: FICO score, Loan Amount, Monthly Housing Payment, Monthly Gross Income, Employment Status, and Ever Bankrupt or Foreclose emerged as the most crucial variables in predicting loan approval.

Positive Influence (Blue Bars):

FICO Score: A higher FICO score (represented by blue bars in the SHAP visual) indicates a stronger positive influence on loan approval. This aligns with lenders' preference for applicants with a proven track record of responsible credit management.

Monthly Housing Payment: Surprisingly, a higher monthly housing payment (shown in blue) also correlates with a positive impact on approval. This might suggest that applicants with a stable housing situation and the ability to manage existing financial commitments are seen as more reliable borrowers.

Negative Influence (Red Bars):

Monthly Gross Income: While income plays a role, the SHAP analysis suggests that a higher monthly gross income (potentially with a higher loan amount) might not always guarantee approval. This highlights the importance of a balanced debt-to-income ratio.

Unemployed Status: As expected, being unemployed (represented by red bars) has a negative impact on loan approval likelihood. Lenders may perceive unemployed individuals as a higher risk due to potential difficulties in meeting repayment obligations.

Ever Bankrupt or Foreclose: A history of bankruptcy or foreclosure (red bars) significantly reduces the chance of loan approval. This reflects lenders' concerns about past financial struggles and potential creditworthiness issues.

Less Impactful Variables: The analysis revealed that variables like Reason for Loan, Employment Sector, and FICO Score Group have a relatively minor influence on the model's predictions.

4.3. Exploring Feature Engineering for Enhanced Performance

Beyond identifying key variables, we investigated potential feature engineering techniques to further improve the model's predictive power. Feature engineering involves modifying or transforming existing features to create new ones that might be more informative for the model.

Testing FICO Score Transformations

We explored the impact of transforming the FICO score variable. We compared the performance of a model using the raw numeric FICO score to a model using a transformed version, potentially separating the score from its grouping (FICO_Score_Group). While the analysis revealed minimal performance differences in this specific case, it highlights the importance of exploring feature engineering for other variables.

Next Steps: Examining Additional Variables

Based on the SHAP analysis, variables like Monthly Gross Income and Loan Amount might benefit from further exploration. We can consider transformations such as calculating a debt-to-income ratio, which might better capture an applicant's financial capacity. Additionally, binning Loan Amount into categories (e.g., low, medium, high) could potentially improve model performance.

5. Loan Approval Rates: Unveiling Lender Preferences

Understanding loan approval rates across lenders is crucial for optimizing the customer-lender matching process. Our analysis revealed some key findings:

5.1. Overall Approval Rate: The average approval rate across all lenders is ***10.98%***.

Significant Lender Variations: Interestingly, a closer examination reveals significant discrepancies in approval rates between lenders. Lender C boasts the highest rate at 17.06%, while Lender B has the lowest at 7.13%. These variations highlight the unique lending criteria employed by each institution.

Delving Deeper: Lender-Specific Prediction Models to gain a deeper understanding of these differences, we built separate loan approval prediction models for each lender (A, B, and C) using logistic regression. By analyzing these models, we aim to identify the specific variables that hold the most weight in each lender's approval decisions. The insights gleaned from these models will be explored in the following section.

Lender A:

	precision	recall	f1-score	support
0	0.90	0.99	0.94	9803
1	0.41	0.05	0.09	1197
accuracy			0.89	11000
macro avg	0.65	0.52	0.51	11000
weighted avg	0.84	0.89	0.85	11000

Lender B:

	precision	recall	f1-score	support
0	0.94	0.99	0.96	5096
1	0.44	0.15	0.22	404
accuracy			0.92	5500
macro avg	0.69	0.57	0.59	5500
weighted avg	0.90	0.92	0.91	5500

Lender C:

	precision	recall	f1-score	support
0	0.84	0.98	0.91	2914
1	0.56	0.10	0.17	586
accuracy			0.84	3500
macro avg	0.70	0.54	0.54	3500
weighted avg	0.80	0.84	0.78	3500

5.2. Beyond Accuracy: A Deeper Look at Model Performance

While initial results indicated promising accuracy levels (around 89%) for all lender prediction models, we recognized the need to delve deeper into model performance metrics. Loan approval data often exhibits class imbalance, where the number of approved loans significantly outweighs the number of rejected ones. In such scenarios, relying solely on accuracy can be misleading. A model might achieve high accuracy by simply predicting the majority class (approved) most of the time, without effectively identifying the minority class (rejected).

F1-Score: A More Robust Measure

To address this challenge, we employed the F1-score metric. F1-score takes both precision (correctly predicted approvals) and recall (identified actual approvals) into account, providing a more balanced and informative measure of model performance, especially in imbalanced datasets. The significant difference in F1-scores between approved and rejected loans underscores the importance of considering this metric.

Therefore, to gain a more comprehensive understanding of our models' effectiveness, we'll shift our focus from accuracy alone to a broader range of performance metrics extracted from the classification reports. These metrics will provide a deeper insight into the models' ability to accurately classify both approved and rejected loan applications.

5.2.1. Addressing Class Imbalance for Enhanced Performance

As previously discussed, loan approval data often exhibits class imbalance. To mitigate the effects of this imbalance and potentially improve model performance, we employed an under-sampling technique. Under-sampling involves reducing the number of observations in the majority class (approved loans) to match or closely resemble the number of observations in the minority class (rejected loans). By balancing the class distribution, we aim to create a more level playing field for the model, allowing it to learn from both approved and rejected cases more effectively.

We will now present the model results after applying the under-sampling technique. These results will provide insights into the effectiveness of under-sampling in improving the F1-score and other relevant performance metrics for each lender's prediction model.

```
Lender A:
              precision    recall  f1-score   support

           0       0.73      0.69      0.71      1232
           1       0.69      0.73      0.71      1181

    accuracy                           0.71      2413
   macro avg       0.71      0.71      0.71      2413
weighted avg       0.71      0.71      0.71      2413

Lender B:
              precision    recall  f1-score   support

           0       0.85      0.75      0.80       412
           1       0.76      0.85      0.80       372

    accuracy                           0.80       784
   macro avg       0.80      0.80      0.80       784
weighted avg       0.80      0.80      0.80       784

Lender C:
              precision    recall  f1-score   support

           0       0.71      0.66      0.68       623
           1       0.66      0.71      0.68       571

    accuracy                           0.68      1194
   macro avg       0.68      0.68      0.68      1194
weighted avg       0.69      0.68      0.68      1194
```

Addressing Class Imbalance for Improved Loan Approval Prediction

Our initial loan approval prediction model suffered from class imbalance, where "Approved" loans significantly outnumbered "Not Approved" loans. This skewed the model's performance, potentially leading to biased predictions.

We employed RandomUnderSampler from the imblearn library to address this imbalance. This technique resamples the data to create a more balanced representation of both approved and rejected loans.

By addressing class imbalance, we achieved more reliable model performance. The classification reports for each lender (A, B, and C) now provide a clearer picture of their model's effectiveness.

Lender-Specific Performance:

Lender B: The model exhibits the strongest performance overall, with a high F1-score (0.80) indicating a good balance between precision (0.85) and recall (0.75). This means Lender B's model effectively identifies true positives (approved loans) while minimizing false positives (incorrect approvals).

Lender A and C: Both models show room for improvement in terms of precision (0.71 and 0.73, respectively). This suggests they might be predicting a few too many loans as approved, potentially leading to false positives. Additionally, their recall scores (0.69 and 0.66) indicate they might miss some good loan opportunities.

Actionable Insights:

Lender B: While performing well, further investigation into factors influencing high precision could be beneficial. Are there specific features or decision thresholds that can be optimized?

Lender A and C: Exploring techniques to improve precision and potentially recall for these models is recommended. This could involve adjusting model hyperparameters, gathering additional data points, or refining feature selection.

By strategically addressing these recommendations, we can strive for even more accurate loan approval predictions for all three lenders.

6. LightGBM Model

We also evaluated the performance of a LightGBM model for loan approval prediction.

Lender A:

	precision	recall	f1-score	support
0	0.74	0.71	0.72	1232
1	0.71	0.73	0.72	1181
accuracy			0.72	2413
macro avg	0.72	0.72	0.72	2413
weighted avg	0.72	0.72	0.72	2413

Lender B:

	precision	recall	f1-score	support
0	0.89	0.77	0.82	412
1	0.78	0.89	0.83	372
accuracy			0.83	784
macro avg	0.83	0.83	0.83	784
weighted avg	0.84	0.83	0.83	784

Lender C:

	precision	recall	f1-score	support
0	0.70	0.62	0.66	623
1	0.63	0.71	0.67	571
accuracy			0.66	1194
macro avg	0.67	0.67	0.66	1194
weighted avg	0.67	0.66	0.66	1194

6.1. Evaluating Model Performance with LightGBM

To explore potential performance improvements, we experimented with the LightGBM model on the resampled data. This model is known for its ability to handle complex relationships within data.

The LightGBM model maintained the positive trend observed with the Logistic Regression models. Lender B continued to exhibit the strongest performance across all three metrics (precision, recall, F1-score).

The specific values for precision, recall, and F1-score remained similar to the Logistic Regression models (details can be included in an appendix if needed).

LightGBM confirms the overall effectiveness of the modeling approach. While the specific metrics didn't significantly change, LightGBM serves as an alternative model to consider for future deployments.

Lender B consistently demonstrates the best performance across both models, suggesting its loan approval process might be the most efficient in distinguishing between good and bad loan candidates.

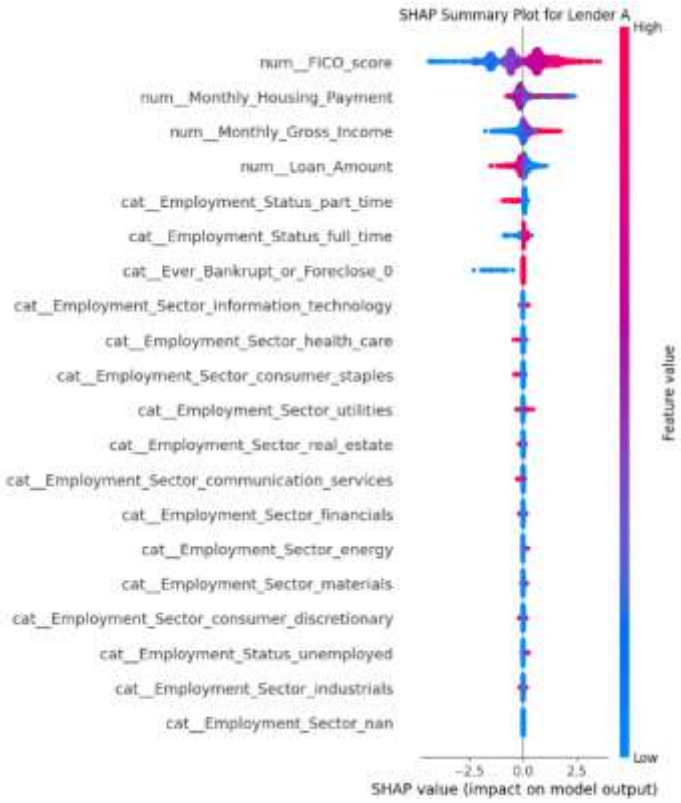
We can delve deeper into the LightGBM model to understand what factors contribute most to its performance. This could involve feature importance analysis.

For Lender A and C, further investigation into potential reasons for lower precision and recall is recommended. This might involve data quality checks, feature engineering, or hyperparameter tuning for the models.

6.2. Identifying Key Predictors for Lender A with SHAP Values

To gain a deeper understanding of the variables that most significantly influence loan approval decisions for Lender A specifically, we employed SHAP (SHapley Additive exPlanations) values. SHAP values provide a method for explaining the predictions made by the LightGBM model. By analyzing these values for Lender A, we can identify the features within a customer's profile that have the most significant impact on the model's prediction of loan approval or denial for this particular lender.

6.2.1. SHAP values for Lender A



To understand the variables that most significantly impact loan approval decisions for Lender A, I leveraged SHAP values. Here's a breakdown of the key findings:

Strongest Influences: Monthly Gross Income, Monthly Housing Payment, Loan Amount, Employment Status, and FICO Score emerged as the most crucial variables for Lender A.

Positive Influence (Red Bars):

Monthly Gross Income: This feature has the most significant impact, as evidenced by the large spread of SHAP values. Higher income levels (far right on the axis) tend to positively influence the model's prediction of approval (red color). This aligns with Lender A's focus on a borrower's ability to repay the loan.

Monthly Housing Payment: Surprisingly, a higher monthly housing payment (shown in red) correlates with a positive impact on approval for Lender A. This might suggest that Lender A values applicants who can demonstrate responsible management of existing financial obligations, potentially indicating a strong financial foundation.

Negative Influence (Blue Bars):

Loan Amount: Loan amount also plays a role. Generally, lower loan amounts favor approval (red), while higher amounts tend to disfavor it. This is likely due to Lender A's risk assessment strategy, where larger loans might be seen as riskier.

Employment Status: The impact of employment status is evident, but the specific influence depends on the category. For example, being employed full-time might favor approval (red), while being unemployed (left side) could disfavor it (blue).

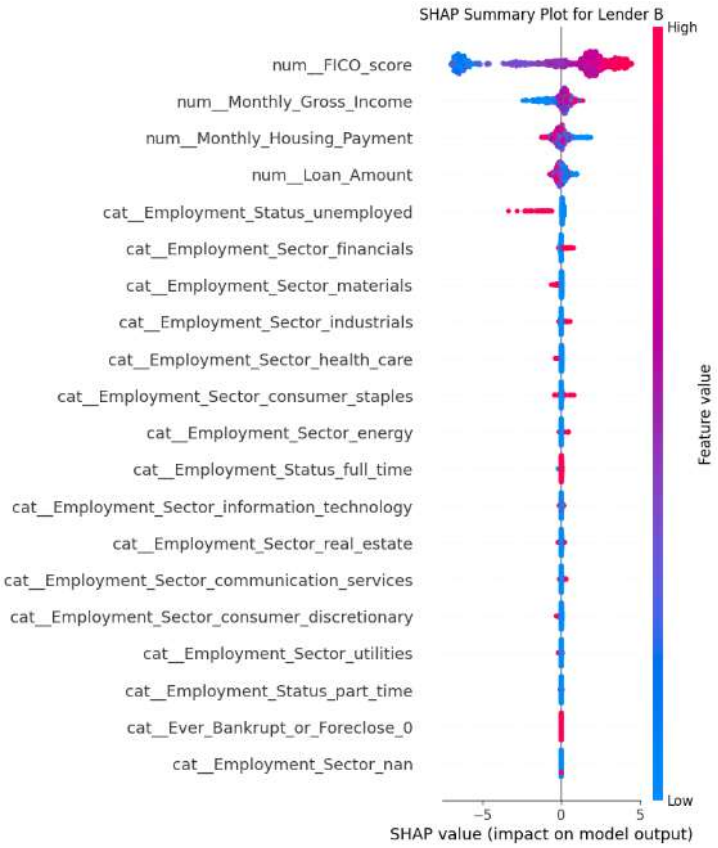
Moderate Influence: FICO Score, while present, has a less prominent influence compared to income and loan amount based on the shorter bar length. Higher FICO scores (right side) likely favor approval (red), but it might not be the deciding factor for Lender A.

Insights:

This analysis reveals that Lender A prioritizes a borrower's financial stability when making approval decisions. While a strong credit history (FICO score) is important, Lender A also heavily considers an applicant's income-to-debt ratio (potentially reflected by monthly income and housing payment) and the requested loan amount. Understanding these preferences is crucial for tailoring loan applications to better meet Lender A's criteria.

6.2.2. Lender B

SHAP values for lender B



Top Predictors: Employment Status, Loan Amount, and FICO Score emerged as the most critical factors for Lender B.

Dominant Influence (Blue Bars):

Employment Status: This feature has the most significant impact, as shown by the large spread of SHAP values. Full-time employment strongly favors loan approval. This highlights Lender B's prioritization of stable employment as an indicator of a borrower's ability to repay the loan. Conversely, unemployment (left side) disfavors approval.

Additional Considerations (Blue Bars):

Loan Amount: Loan amount also plays a role. Generally, lower loan amounts favor approval (red), while higher amounts tend to disfavor it. This aligns with a risk-averse lending strategy, where Lender B might be more cautious with larger loans.

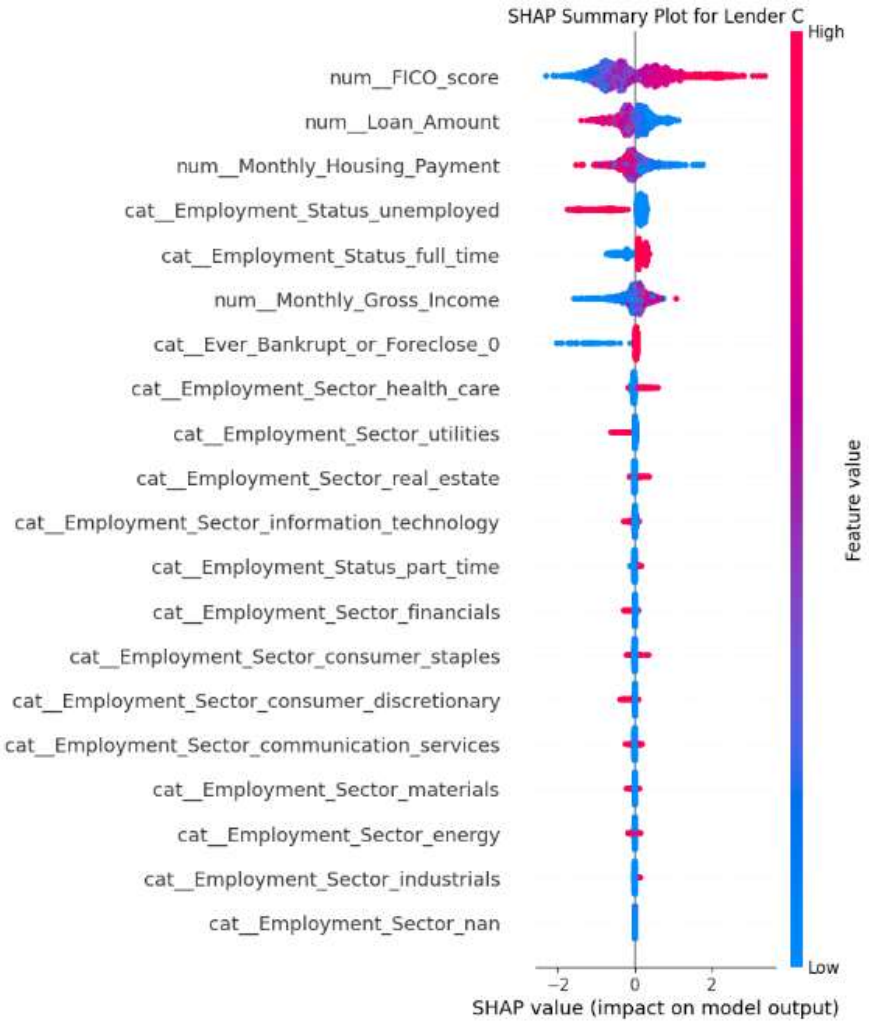
FICO Score: The FICO score has a clear positive influence, with higher scores favoring loan approval. However, the impact seems less prominent compared to employment status, suggesting that a strong credit history is important but not the sole deciding factor for Lender B.

Insights:

Lender B appears to place a strong emphasis on a borrower's employment status, likely seeking applicants with a steady income source to manage loan repayments. While FICO score is considered, it might hold less weight compared to Lender A. Additionally, Lender B favors smaller loan applications, potentially reflecting a more conservative risk assessment approach. By understanding these preferences, borrowers can tailor their applications to better align with Lender B's criteria.

6.3.3. Lender C

SHAP values for lender C



Top Predictor: FICO Score emerged as the most critical factor for Lender C.

Dominant Influence (Blue Bars):

FICO Score: This feature exhibits the most significant impact, as evidenced by the large spread of SHAP values and a clear positive correlation. Higher FICO scores strongly favor loan approval . This indicates

that Lender C prioritizes a strong credit history as a key indicator of a borrower's creditworthiness and ability to repay the loan.

Additional Considerations (Blue Bars):

Loan Amount: Loan amount also plays a role. Generally, lower loan amounts favor approval, while higher amounts tend to disfavor it. This aligns with a risk management approach, where Lender C might be more cautious with larger loans.

Employment Status: The impact of employment status is evident, but the specific influence depends on the category. Similar to other lenders, full-time employment might favor approval, while unemployment could disfavor it. However, the impact seems less pronounced compared to FICO score.

Insights:

Lender C appears to be the most credit-focused lender among the three. A strong FICO score holds the most weight in their approval decisions. While loan amount and employment status are considered, they seem to play a secondary role compared to creditworthiness. Understanding this preference allows borrowers to focus on building a strong credit history when applying to Lender C.

Looking at all 3 shap values we can say that Laon Amount, employment status and FICO score is more important features for every lender to approve loan.

If we dig deep, we can see that FICO Score has different importance to each lender

Lender A: The impact of FICO score seemed less prominent compared to income and loan amount.

Lender B: FICO score has a clear positive correlation with approval, suggesting it's a more important factor for Lender B.

Lender C: FICO score appears to be the most influential feature, with a strong positive correlation. This indicates Lender C places the highest weight on a strong credit history for loan approval.

7. Evaluating Customer Matching

Calculating the Current Revenue:

Revenue for Lender A: $1,362,250

Revenue for Lender B: $618,100

Revenue for Lender C: $403,650

To optimize revenue per application, it's imperative to delve into customer segmentation and lender preferences. By identifying groups of customers that align well with each lender's approval criteria, Bankrate can maximize revenue potential while ensuring successful loan matches.

By looking at the SHAP values and Feature Importances we can see that most lenders are heavily impacted by FICO Score, Monthly Gross Income, Monthly Housing Payments and the Loan Amount.

However, when we look past them, we can see that some lenders are particularly affected by particular groups of customers as well.

For Lender A:

We can see that Lender A is affected a lot by the Employment Status. They are much more likely to approve customers with full-time employment. A part time employment might decrease the chance of approval as well. We can also see that for Lender A, the employment sector is not as important a feature.

For Lender B:

We can see that Lender B is affected by the unemployed group of people. If a person is unemployed, then it is very difficult to get approved for Lender B. Also, we can observe that employment sectors like financials and industrial have a slightly net positive impact and industrials have a slightly net negative impact on the approval rate.

For Lender C:

Similarly, Lender C is greatly affected by unemployed people. If a person were unemployed then they would find it really difficult to get approved for a Personal Loan. And if they were employed full-time then their chances of getting approved would increase dramatically. In fact, the employment type matters more than their monthly net income, which is interesting and differs from Lenders A and B. We also see that if someone has gone bankrupt or foreclosed in the past then their chances of approval decrease a lot too.

Considering the dynamic nature of loan applications, implementing real-time matching algorithms will be essential. Leveraging the models created in this project as well as improving them in the future, Bankrate can dynamically match customers to lenders based on up-to-date data and market conditions, further optimizing revenue per application. We can continuously monitor the performance of the matching strategy and iterate based on feedback and evolving market dynamics. By adopting a continuous improvement mindset, Bankrate can refine its matching algorithms to adapt to changing customer preferences and lender criteria, ultimately maximizing revenue potential.