“Should This Loan be Approved or Not?”

[Aneesh Krishna, Sathwik Veeraiah Udgimath, Ashray Jaya Mani, Jatin Kamalkishor Parihar]
December 23, 2023

1. **Problem Importance**

   Significance: In our loan approval project, our main goal is to predict the likelihood of a small or medium-sized business not being able to pay back the loan. This is crucial for our bank because if these businesses can't repay, banks might face financial losses. To achieve this, we're using a large set of data from past loan applications. We're looking for patterns or characteristics in these applications that can tell us if a loan is more or less likely to end in a situation where the business can't repay. This predictive model will help us make smarter decisions about which loans to approve, reducing the risk of financial problems for the banks.

   Consequences of Incorrect Decisions: Incorrect decisions in loan approvals can have profound consequences. Approving loans to high-risk applicants may lead to increased default rates, financial losses, and potential systemic impacts. On the other hand, unnecessary rejections can limit credit access for deserving individuals, hindering economic development.

2. **Problem Analysis**

   Methodology: The analysis employed a two-fold approach, utilizing both Logistic Regression and Random Forest models. Logistic Regression provided a baseline, while the Random Forest model demonstrated superior predictive power.

   Features and Preprocessing:

   - **Selected Features:**

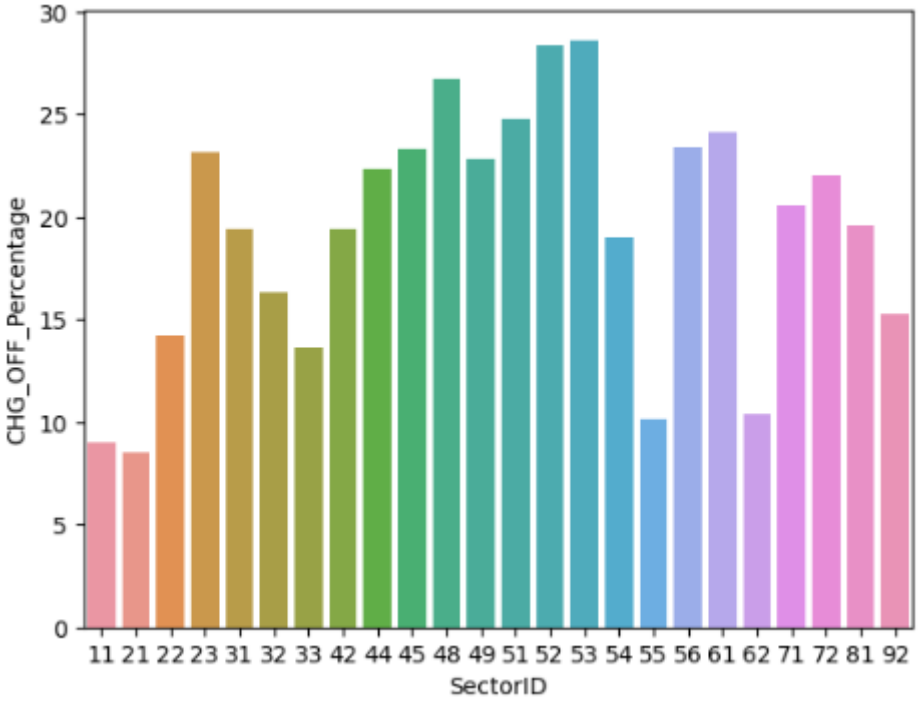
Features	Description
RECESSION	1 if loan is active during Great Recession, 0 otherwise
REALESTATE	1 if loan is backed by real estate, 0 otherwise
PORTION	Proportion of gross amount guaranteed by SBA
DISBURSEMENTGROSS	Amount disbursed
RETAINEDJOB	Number of jobs retained
CREATEJOB	Number of jobs created
NEWEXIST	1 is Existing business, 2 is New business
NOEMP	Number of business employees
TERM	Loan term in months
URBANRURAL	1 if Urban, 2 if rural

   - **Preprocessing Steps:**
     Handled missing values: Dropped rows or imputed values based on the nature of missing data.
     Scaled numeric features: StandardScaler was applied to ensure uniform scales.
     Derived features: 'DEFAULT' was created to represent loan outcomes.

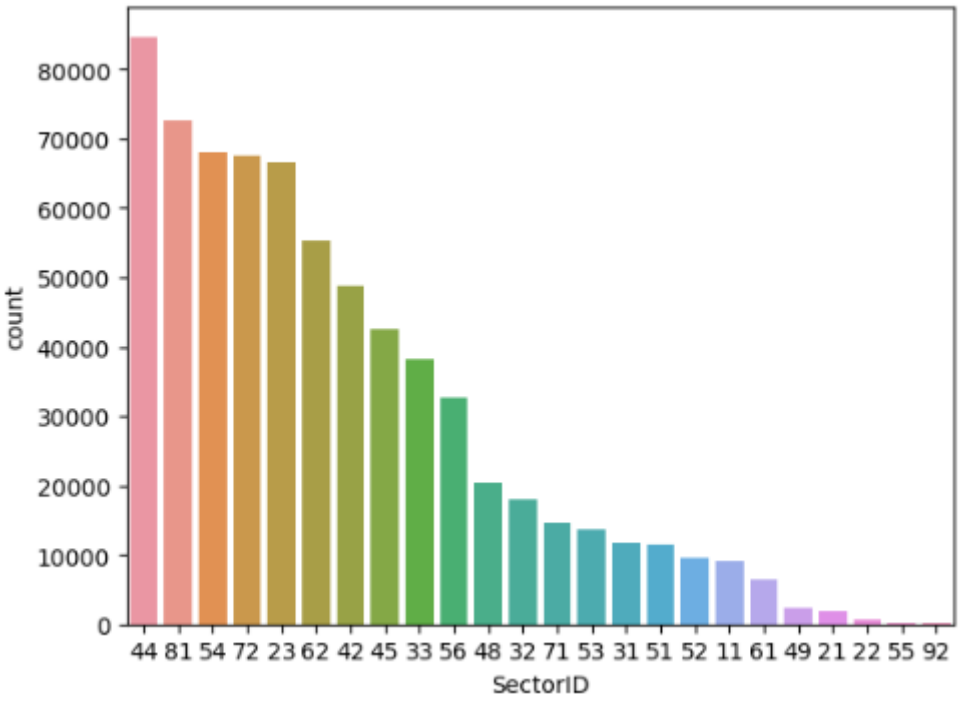
Features	Description
RECESSION	Filtering out from December 2007 to June 2009 based on DISBURSEMENTDATE
PORTION	SBA_APPROVED_AMOUNT/GROSS_APPROVED_AMOUNT
REALESTATE	If Term is > = to 240 then 1, else Term < 240 is 0
DEFAULT	1 if MIS_Status is CHGOFF, 0 if MIS_Status is PIF

3. Findings and Visualization:

Exploratory Data Analysis



Based on the graph, it is evident that sectors 52 and 53 have a higher likelihood of loan default, while sectors 11 and 21 demonstrate a higher rate of customers successfully repaying their loan amounts in full.



This graph provides an overview of the number of loan applications within each specific sector.

Models

The Random Forest Classifier outperforms Logistic Regression in terms of accuracy and overall predictive performance, as evidenced by higher precision, recall, and F1-scores for both classes.

Logistic Regression has a higher precision for non-events, but Random Forest has a better overall balance between precision and recall for both classes.

Random Forest shows a significant improvement in identifying actual events compared to Logistic Regression.

Depending on the specific goals and considerations, the Random Forest Classifier may be a better choice for this classification task.

Model	Accuracy	True Negative	False Positive	False Negative	True Positive
Logistic Regression	0.83	143681	4189	25602	6123
Random Forest Classifier	0.93	142476	5394	7780	23945

Correlation Matrix Heatmap

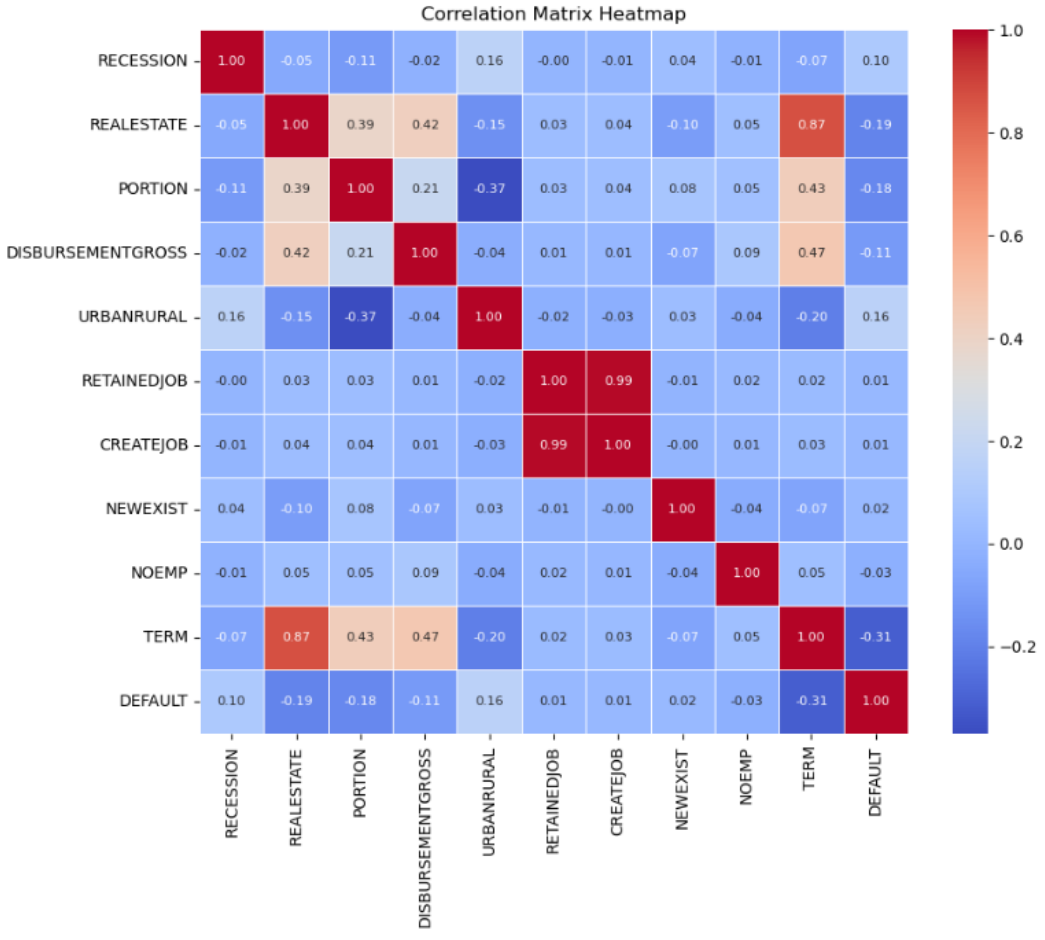
Key Relationships:

Positive correlations with 'REALESTATE,' 'DISBURSEMENTGROSS,' and 'PORTION.'

Negative correlations with 'RETAINEDJOB' and 'URBANRURAL.'

Insights:

Loans associated with real estate, higher disbursement amounts, and specific portions of the approved amount have a higher likelihood of default.



4. Conclusion

Exploratory Data Analysis and machine learning models, particularly the Random Forest, offer nuanced insights into loan repayment patterns. Implementing these insights in lending decisions enhances the ability to assess creditworthiness accurately. This data-driven approach contributes to responsible lending practices, minimizing risks and promoting economic stability.

References:

[1] Should This Loan be Approved or Not? from Kaggle.
https://www.kaggle.com/datasets/mirbektoktogaraev/should-this-loan-be-approved-or-denied