DataDriven
PORTFOLIO SOLUTIONS

Stage B report: Data Ingestion and EDA

In this report, we summarize the progress made in the Data Understanding/Data Preparation step of our project. This step involved ingesting, cleaning, and conducting exploratory data analysis (EDA) on the provided peer-lending data to gain insights and identify potential issues. Our initial findings and insights from the EDA are outlined in this working paper, additional technical details available in the attached code file (stepB.py).

As previously agreed, we are pleased to present to you the distribution and expected return for the different loan grades in Appendix C. This information is vital in understanding the potential risks and returns associated with the investment. It is important to note that this data is based on historical trends and does not guarantee future performance.

Documenting Completed Work Assignments-

1. **Data Consistency and Consolidation:** Our dataset includes loans granted in 2016, organized into 8 files, with 4 files for each year containing snapshots of the same loans for 2018 and 2019, classified according to the quarter in which the loan was taken. To ensure data consistency, we consolidated the files by year, allowing us to:
   a. Confirm the presence of identical loans across years and quarters.
   b. Checked feature consistency across years, except for one unique to 2018.
   c. Perform a leak test- We calculated the percentage change in attribute values for each loan identifier between 2018 and 2019 to identify potential data leakage.

2. **Selecting Relevant Loans:** The raw data contains 434407 instances. We removed instances with loan statuses other than "Charged Off" and "Fully Paid" to ensure completed loan outcomes, resulting in 356,988 instances. A graph depicting the loan status count can be found in Appendix A.

3. **Validating Data Types:** Data type conversion and checking for appropriate data types were performed on columns of a dataset for accurate and efficient analysis.

4. **Target variable exploration:** We calculated two target variables for our analysis: a continuous variable representing the yield for each loan, and a binary variable indicating whether the loan yield is above 2%. At this point, we will explore both variables and determine which one we will use to develop our model within 3-4 weeks.

5. **Feature selection:** The raw data contains the 150 features. Our approach to preliminary selecting features was guided by the following considerations (Refer to the attached Excel file's Column Status sheet for details on each column's status.):
   a. Features exhibiting leakage-We performed a thorough analysis and removed 42 features that exhibited changes between 2018 and 2019, and identified additional features that could

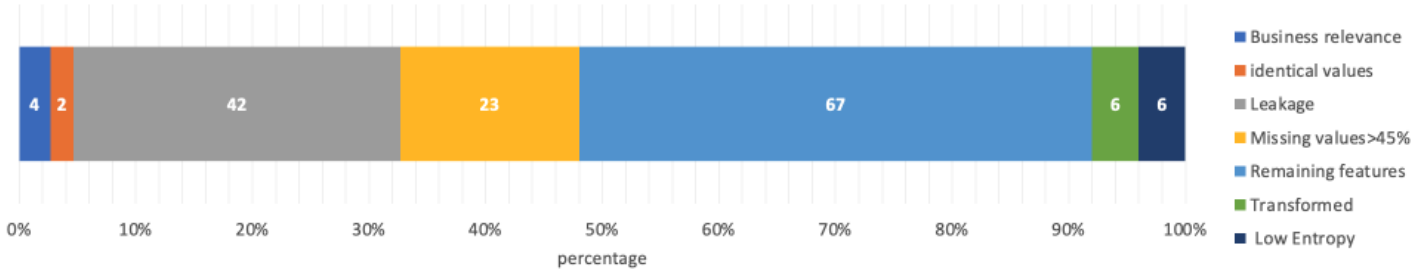
potentially cause leakage. This ensures that our model is not trained on features that will not be available at the time of prediction, thus improving its reliability.

b. Missing values (NA's) per feature- To prevent inaccurate data imputation and biased models, we removed 23 features with more than 45% missing values, while considering the missing value percentage of subsequent features (13.2%). A heatmap in Appendix B illustrates the missing values distribution, it is worth noting that there were not many missing values in the dataset.

c. Features business relevant- Four features that did not align with our business understanding were identified and removed from the dataset, as they could add noise and negatively impact the model's performance-Appendix B.

d. The distribution of features values-Our evaluation of the feature values distribution relied on various techniques. These included removing features with identical values for all observations and analyzing the feature entropy and low frequency values. We made decisions on whether to remove or modify the features based on their relevance to the problem at hand, taking into account factors such as the percentage of instances for each category/distribution according to the target variable (Column Transformation and low frequency values sheets).

Feature Selection Distribution-



Ongoing feature selection involves examining and making decisions on each feature's relevance. The process should finish in three weeks with the model construction phase.Appendix -C:H

6. **Loan Application Analysis:** As previously stated, we considered the potential impact of low frequency values on the accuracy and reliability of our analysis. Our examination revealed minimal occurrences of joint loan applications, which were removed from our analysis to mitigate the risk of bias and account for possible distinctions between individual and joint loan applications.

Loan Status by Application Type-

loan_status	Charged Off	Fully Paid	Total
application_type			
Individual	98.05%	98.23%	98.19%
Joint App	1.95%	1.77%	1.81%

7. **Missing Data Handling:** We conducted a comprehensive analysis of the missing data in the dataset, identifying features with a low percentage of missing values (<0.06) for which instances with missing

values were removed. For features with higher percentages, we employed various techniques such as linear regression model and grouping. We have currently imputed some of the missing values using statistical measures and anticipate that smarter imputation methods will take two weeks to complete. A detailed explanation of the methods used to impute the missing data, along with relevant graphs for each column, is available in the Handling Missing Values sheet.

8. Performing correlation analysis is crucial for developing predictive models and feature selection. We'll analyze the relationship between our target variable and the features to identify the most important variables. Correlation coefficients will measure the strength and direction of the linear relationship between variables. This analysis helps select the most important features for predictive models and identify potential multicollinearity issues. See Appendix-I.

Potential Pitfalls:

- Incorrect data provided by the borrower can lead to biased results.
- The presence of outliers in many of the features require an estimated treatment time of two weeks.
- Improperly handling missing values can introduce bias and lead to incorrect results.
- Incorrectly correcting or removing outliers can also introduce bias and lead to incorrect results.
- Using the wrong type of transformation can lead to incorrect or meaningless results.
- Incorrectly encoding categorical variables can lead to incorrect assumptions about their relationship with the target variable.
- Creating features that are highly correlated with existing features can lead to overfitting and negatively impact model performance.

Our team will spend two weeks continuing analyzing, cleaning, and addressing potential issues in the data. We will present our findings and recommendations in a detailed report. Appendix J outlines our work steps.

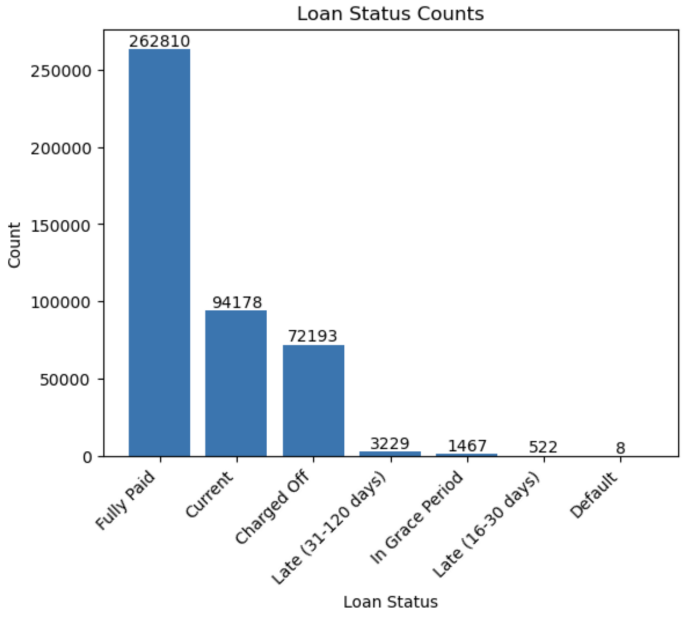
Best regards,

The DataDriven Portfolio Solutions Management Team

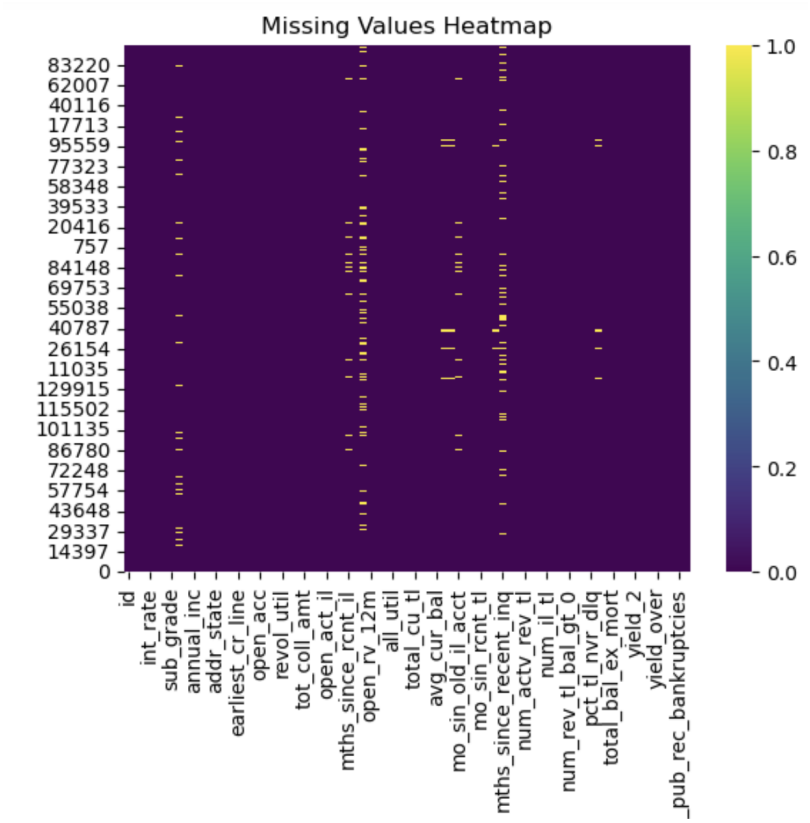
DataDriven
PORTFOLIO SOLUTIONS

Appendices

Appendix A:

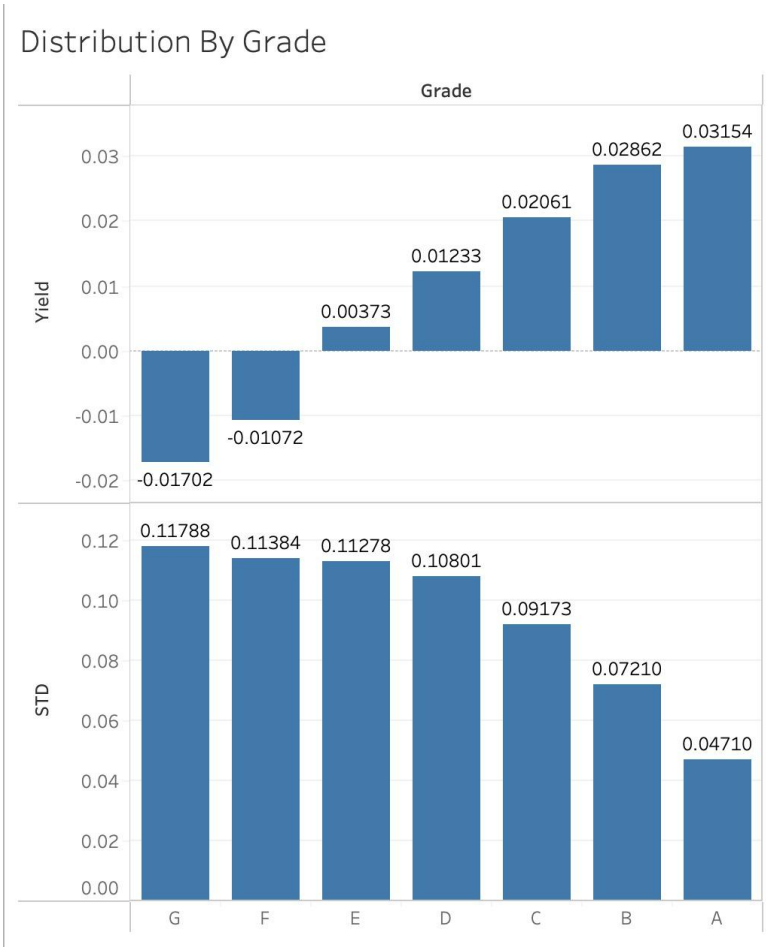


Appendix B:

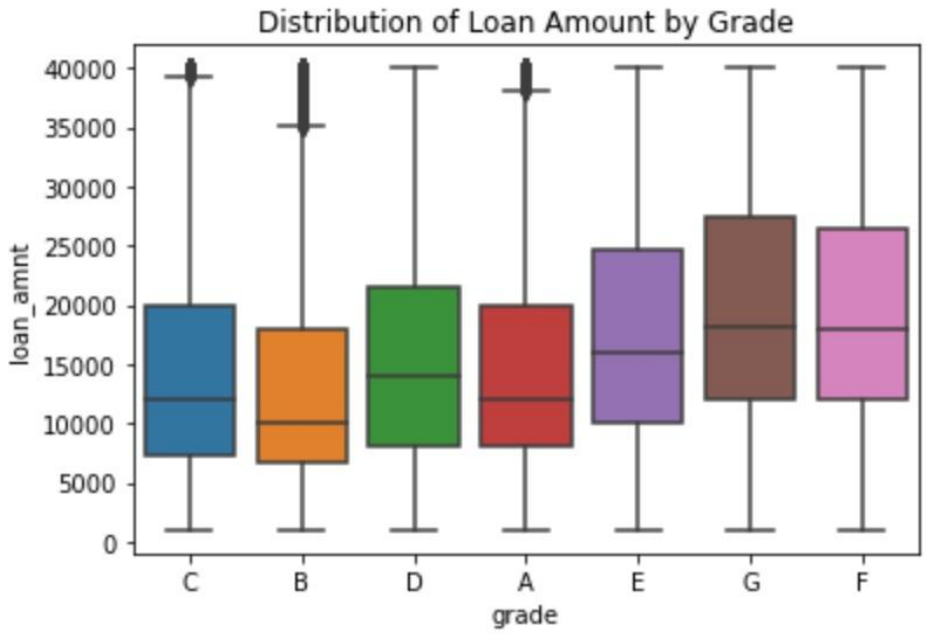


DataDriven
PORTFOLIO SOLUTIONS

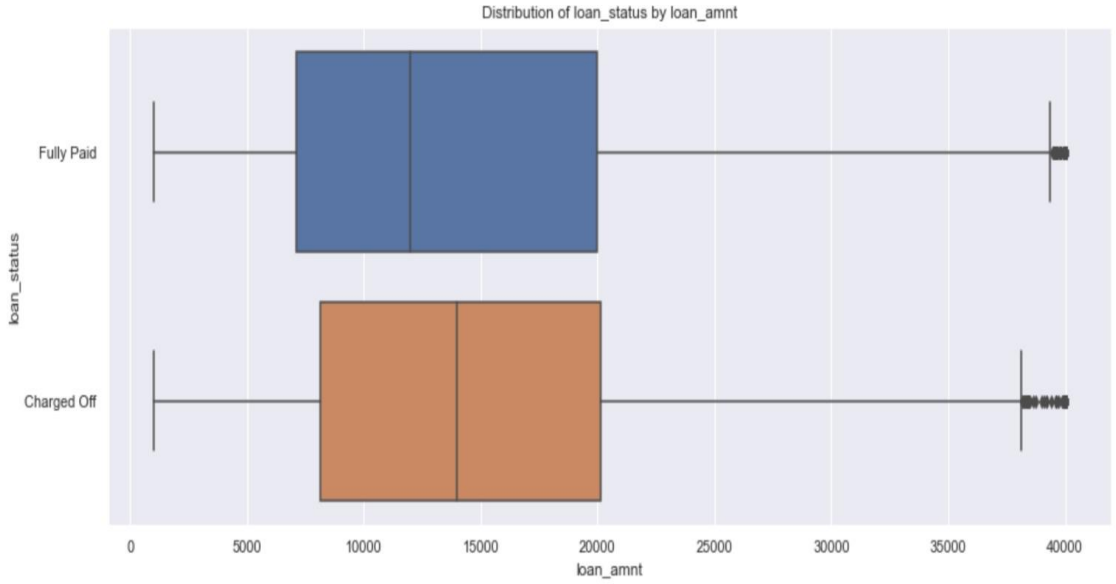
Appendix C:



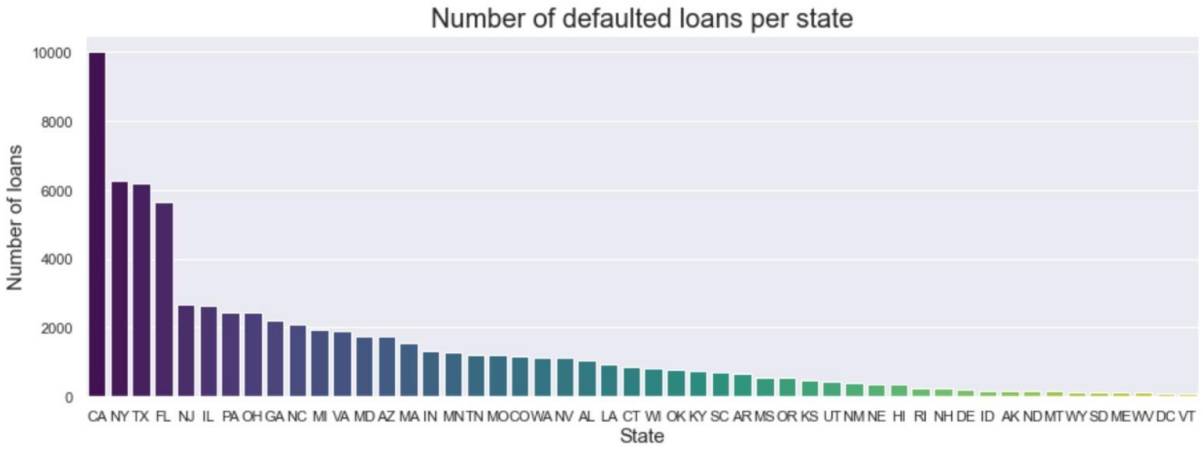
Appendix D:

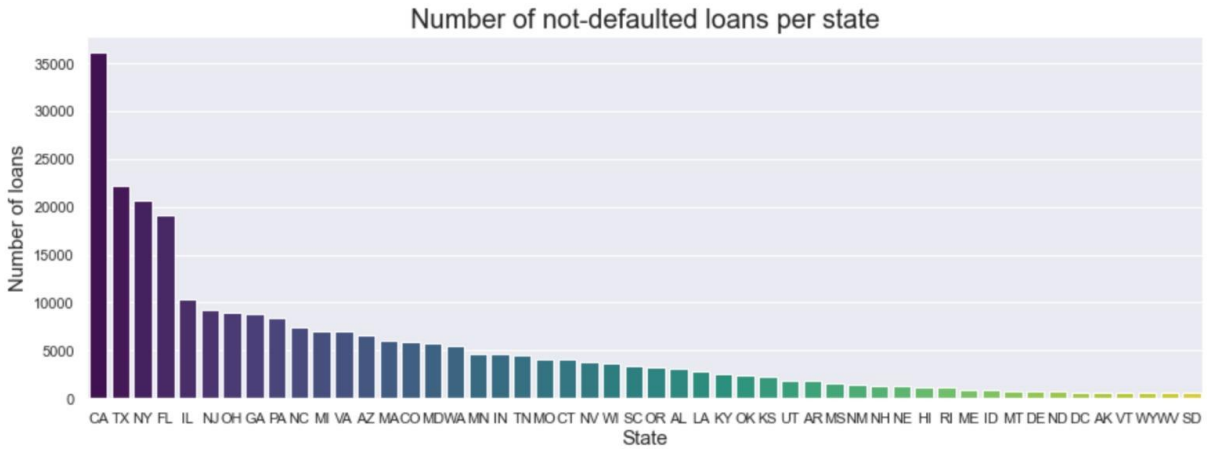


Appendix E:



Appendix F:

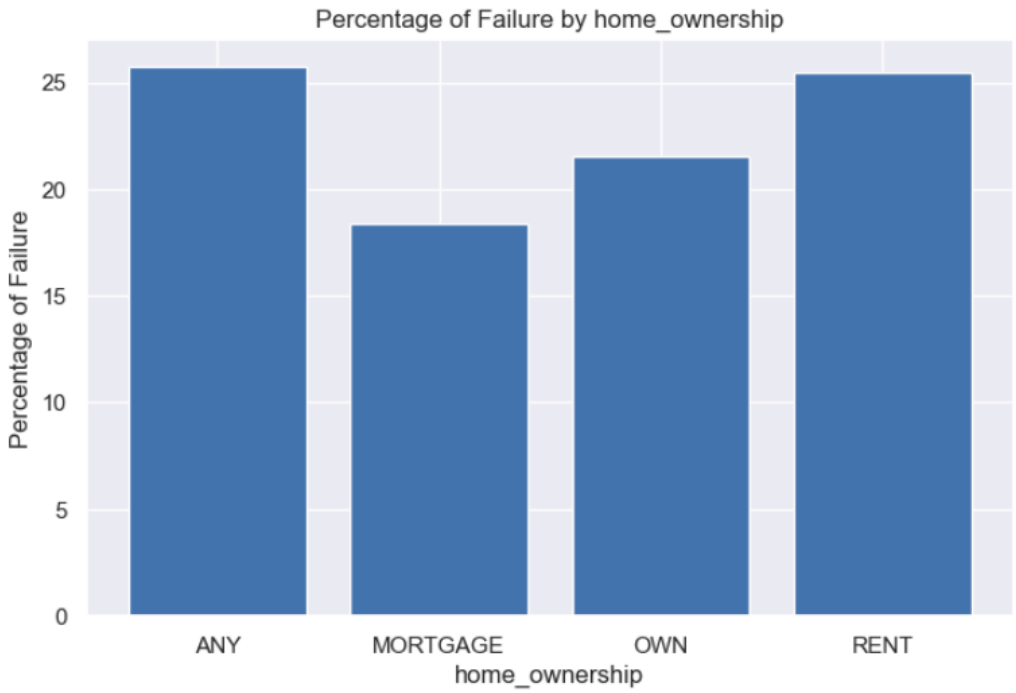




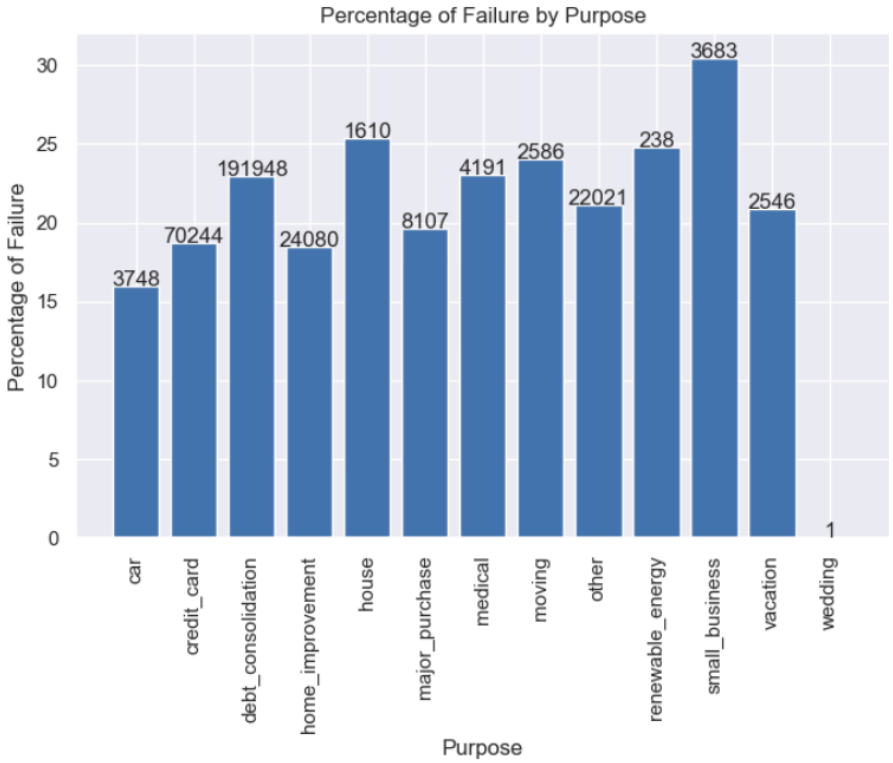
-

DataDriven
PORTFOLIO SOLUTIONS

Appendix G:



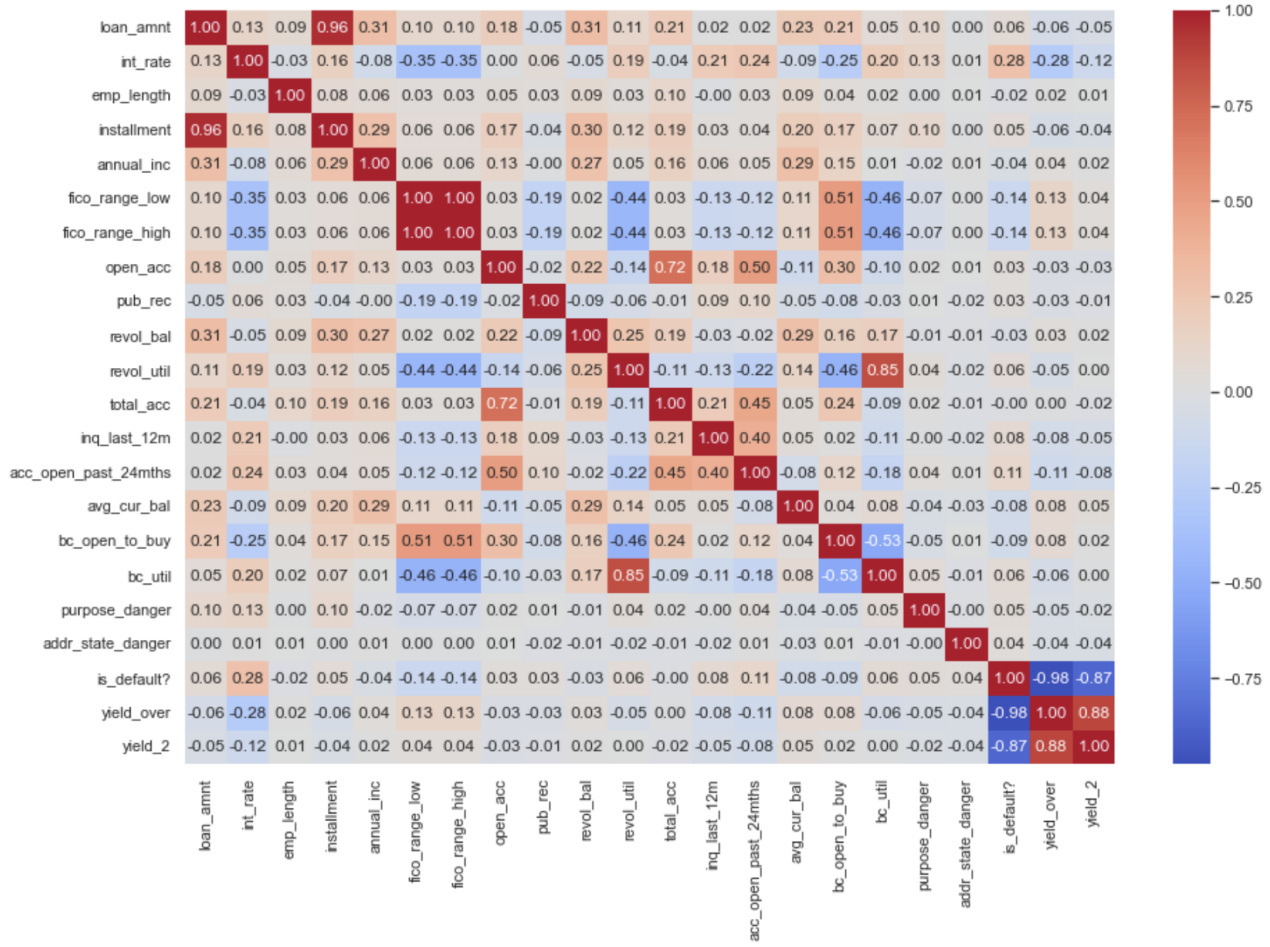
Appendix H:



DataDriven
PORTFOLIO SOLUTIONS

Appendix I:

→



DataDriven
PORTFOLIO SOLUTIONS

Appendix J:

Time-line

Task	Status	Description	time estimation
2	In-progress	EDA - process of analyzing and understanding the characteristics of a dataset through visual and numerical methods.	2-3 week
2.1	In-progress	Handling NA	2-3 week
2.2	Not-Started	Handling outliers	2-3 week
2.3	In-progress	Visualizing features in the EDA part (distribution , NA , outliers)	2-3 week
2.4	Done	Handling unique values	-
2.5	Done	choosing the target variable	-
2.6	Done	Handling leakage data	-
2.7	In-progress	Features business relevant	2-3 week
3	In-progress	Data preparations for the model	4-5 weeks
4	Not-Started	Selecting the models	4-5 weeks
5	Not-Started	Building the models + evaluation	5-6 weeks
6	Done	Calculate the return	5-6 weeks
7	Not-Started	investigate and determine the business significance of the findings	7-8 weeks
8	Not-Started	provide the best answers to the question	7-8 weeks