Stage C Report: Data Preparation and Model Selection

This report presents the progress made in the Data Preparation and Modeling step of our project, building on insights from the previous Data Ingestion and EDA report. We outline the data processing steps, considerations for selecting the appropriate model, and approach for evaluating multiple models. Findings are detailed in the report, with technical details available in the attached code file (stepC.py).

Documenting Completed Work Assignments-

Our approach challenges the conventional data division into two types, acknowledging its limitations in capturing real-world complexity. Instead, we adopt a nuanced classification of seven types: useless, nominal, binary, serial, count, time, and interval. This empowers us to understand feature diversity and apply specialized techniques accordingly, capturing vital nuances and avoiding information loss. In this step, our goal is to optimize future feature selection. We exclude useless columns and focus on rigorously tested features. By preparing the remaining features for direct model integration, we minimize future data preparation efforts. This enables us to prioritize model refinement and insightful analysis, freeing us from laborious data preprocessing tasks.

1. **Progress in the data cleaning process as indicated in the previous report:**
   a. **Selecting Relevant Loans:** The raw data contains 434,407 instances that decrease in the previous step to 328,956 instances, with 99,404 instances removed due to loan status other than "Charged Off" and "Fully Paid," and 6,047 joint loan applications removed to mitigate bias (Refer to stage-B report for more information). At this stage, we removed 42 instances from our analysis due to unusual changes observed in the int_rate and installment columns between 2018 and 2019 snapshots. Considering the unlikely possibility of actual changes in these columns, we attributed these discrepancies to potential human error, which could potentially compromise the reliability and accuracy of our analysis. Therefore, we ultimately analyzed 328,914 instances.
   b. **Feature selection:** In this step, we conducted further investigation on the initial selection of features from the raw data containing 151 features, based on the following considerations (See attached Excel file for column feature updates):
      i. Features exhibiting leakage-We removed 43 features that identified as potential leakage features (refer to stage-B report for further details).
      ii. Missing values (NA's) per feature- We removed 23 features with more than 45% missing values (refer to stage-B report for further details).
      iii. Features business relevant- We removed 7 features that didn't align with our business understanding (see stage-B report for further details).
      iv. The distribution of features values- In addition to our previous assessment of entropy and low frequency values, we performed tests for standard deviation and high frequency values to assess attribute value distribution in this stage due this we removed 8 features. The reason we performed these additional tests was to identify highly scattered/not scattered values and values with a high frequency in one value, respectively. We made decisions on whether to remove or modify features

based on their relevance to the problem and the distribution by target variable (refer to Appendix A for further details for each column and the report of Step B).

2. **Handling missing data:** We addressed missing data by employing methods such as linear regression and clustering, and used statistical measures to impute some of the missing values. To avoid compromising the true distribution of each feature, we used two completion methods for each feature. Details and corresponding graphs for each column can be found in Appendix B. We plan to evaluate the effectiveness of these methods by using F1, r^2, and other metrics on our model in the following weeks.

3. **Selection of models:** In the previous step, we computed two target variables: loan yield (continuous) and yield above 2% (binary). To understand p2p loan returns comprehensively, we selected a regression model for quantitative yield estimation and a classification model to identify loans meeting the target return threshold (2%). These models will evaluate the profitability and risk of integrating p2p loans into investment portfolios. Over the next two weeks, we will test various models: K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machines, and Logistic Regression for classification, and Linear Regression, Polynomial Regression, Decision Tree Regression, and Random Forest Regression for regression. In 4 weeks, following the performance evaluation outlined in the last paragraph of this report, we will select the optimal model for implementation.

4. **Handling Outliers:** We examined the distribution of variables to identify and handle outliers. Histogram and QQ plots were used to understand the distribution of each feature. For normally distributed variables, we employed the IQR method and considered values within three standard deviations from the mean as non-outliers. For non-normal distributions, we explored techniques such as log transformation and features and transformation. (refer to Appendix E for further details for each column)

5. **Correlation between features:** We considered data type, distribution, and variable type to choose the appropriate correlation method. For continuous features, the assumption of normal distribution and linearity associated with Pearson correlation coefficient was evaluated. However, as our numeric continuous features did not follow a normal distribution, we opted for Spearman's rank correlation coefficient, a non-parametric test. A correlation heatmap was created to visualize correlations, a threshold of 0.9 correlation score which indicates a strong linear relationship between the variables was chosen to remove correlated variables and scatter plots helped examine the highly correlated variables. To manage multicollinearity, two variables were removed and a new feature was created from correlated features. If any more features need to be removed, a lower threshold will be chosen. For categorical variables, we used the chi-square test to assess associations and uncover meaningful patterns and dependencies. (refer to Appendix F for further details for each column)

6. **Feature Engineering:** We aimed to improve the predictive power of the model by generating new features through various transformations applied to existing features. Our goal was to identify informative features while preventing overfitting. The need for this process arose due to low frequency of values, outliers, business understanding, high correlations between variables, and handling categorical variables etc.. The Phase B report details the entire process, with a list of new features created in Appendix C. In total, we generated 12 new features.

7. **Handling Categorical Features:** Continuing from the previous step, we addressed the issue of handling categorical variables in our model. To simplify the model, we adopted a general approach that involved reducing the dimensions of the categorical variables. We used several techniques such as category transformation based on their relevance to the target variable, one-hot encoding, and hashing. For a detailed explanation of each variable's treatment and corresponding graphs, please refer to Appendix D and the report of Step B.

8. **Standardization:** We applied min-max normalization to standardize the parameters and account for handled outliers. This method scales parameter values between 0 and 1, maintaining the relative relationships between data points. It ensures equal treatment and consistent scaling of all parameters, which is crucial for machine learning algorithms. We will continue testing and may make adjustments as needed due to sensitivity to outliers for this method.

Potential Pitfalls:

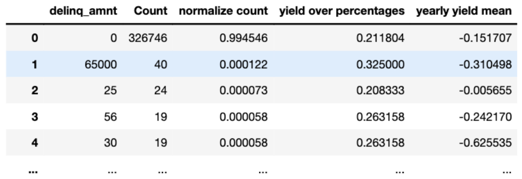
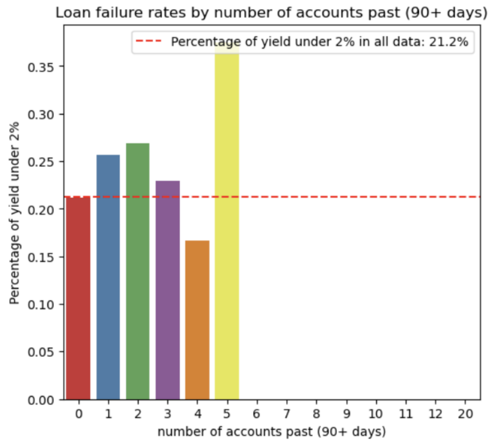
- Incorrect data provided by the borrower can lead to biased results.
- Inaccurate imputation methods that may introduce bias or distort the true distribution of features.
- Insufficient evaluation of missing value completion methods.
- Limited evaluation of models using relevant metrics and benchmarks.
- Inappropriate assumptions about variable distributions leading to ineffective outlier handling.
- Limited exploration of alternative outlier detection and treatment techniques.
- Insufficient assessment of multicollinearity's impact on model stability and interpretability.
- Overcomplicating the model with excessive new features without proper justification.
- Failure to address potential overfitting issues caused by new feature introduction.
- Incorrect type of transformation leading to incorrect or meaningless results.
- Incorrect encoding of categorical variables can lead to incorrect assumptions about their relationship with the target variable.
- Arbitrary dimensionality reduction methods that may result in information loss.
- Neglecting the impact of outliers on the chosen standardization method's effectiveness.

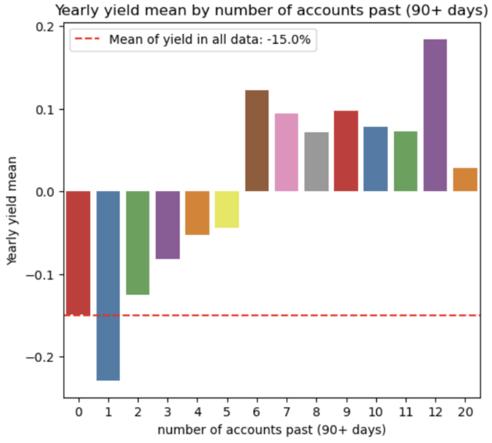
Our next steps involve training, optimizing, and evaluating the model's performance using metrics like accuracy, precision, recall, and F1-score. We will fine-tune the model through techniques like hyperparameter tuning and cross-validation to optimize performance and assess robustness. Through iterative improvement, we will refine the model by exploring different approaches, algorithms, or parameters. We will compare the performance of the regression and classification models, considering business and performance perspectives. Additionally, we will compare our model with the existing loan grade model, using methods like forecasting the probability of returns above 2% and divide them into 7 categories (A-G), comparing the percentage of loans predicted correctly in each grade. We will use Python to perform these operations. Anticipated timeframe for clear results is 5 weeks.

Best regards,
The DataDriven Portfolio Solutions Management Team

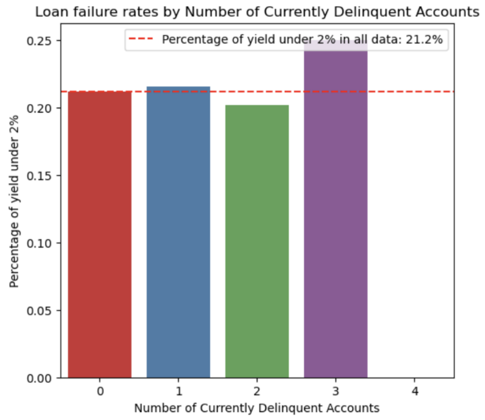
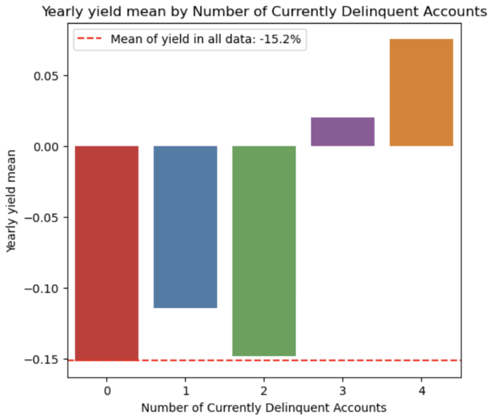
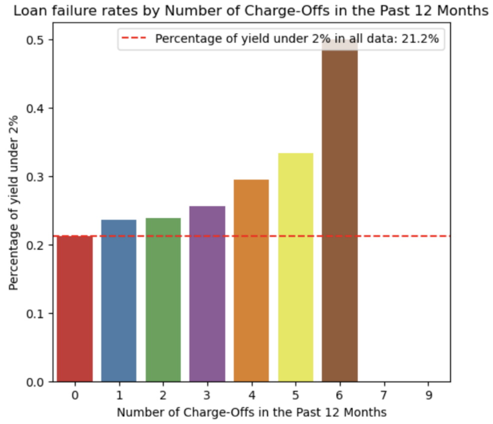
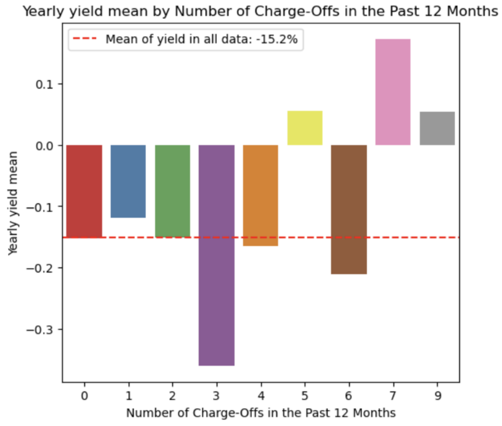
Appendices

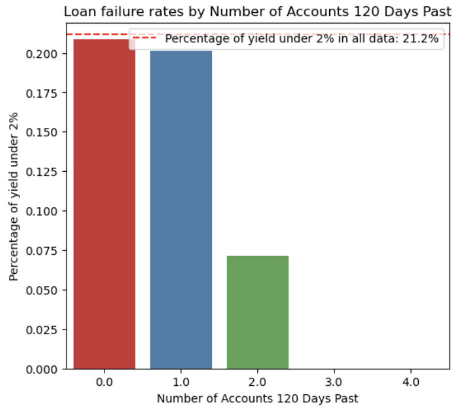
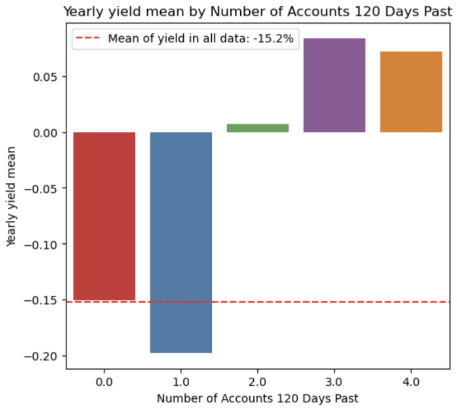
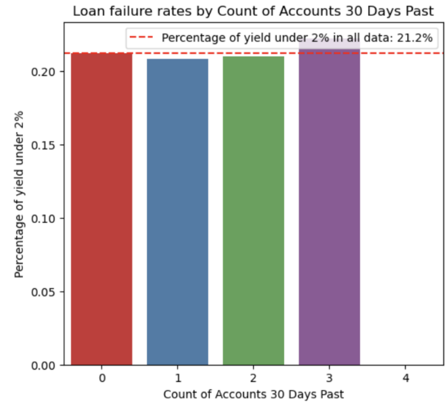
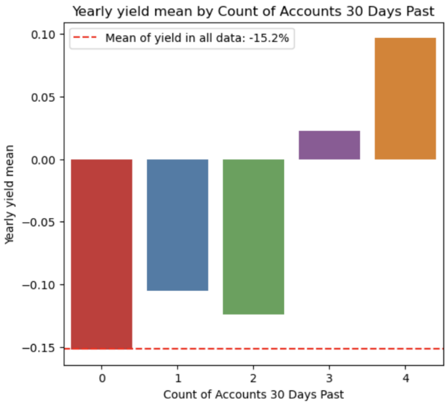
Appendix A:

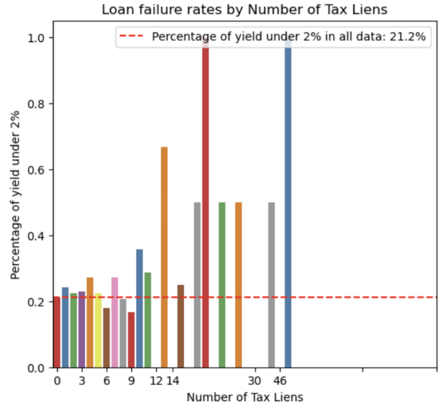
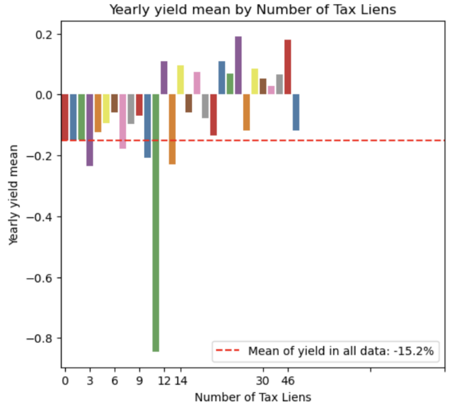
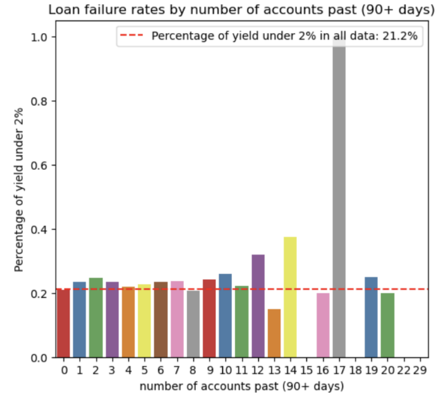
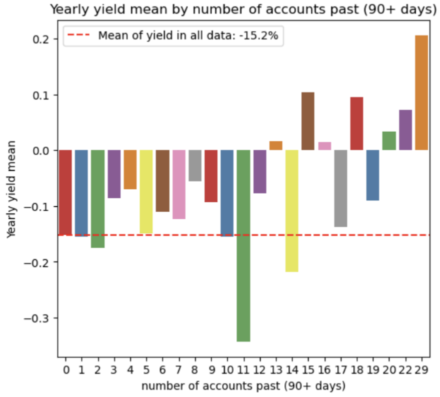
<table>
<tr><th>Column Name</th><th>Treatment</th><th></th></tr>
<tr>
<td><b>delinq_amnt</b></td>
<td>This variable contains 99.45% of the data in one category, where both the failure and return rates are close to the overall average. Therefore, we will remove this variable and thus prevent potential biases or inaccuracies in our analysis that could arise from including a variable with minimal variation and limited impact on the target variable.</td>
<td>
<table>
<tr><th></th><th>delinq_amnt</th><th>Count</th><th>normalize count</th><th>yield over percentages</th><th>yearly yield mean</th></tr>
<tr><td>0</td><td>0</td><td>326746</td><td>0.994546</td><td>0.211804</td><td>-0.151707</td></tr>
<tr><td>1</td><td>65000</td><td>40</td><td>0.000122</td><td>0.325000</td><td>-0.310498</td></tr>
<tr><td>2</td><td>25</td><td>24</td><td>0.000073</td><td>0.208333</td><td>-0.005655</td></tr>
<tr><td>3</td><td>56</td><td>19</td><td>0.000058</td><td>0.263158</td><td>-0.242170</td></tr>
<tr><td>4</td><td>30</td><td>19</td><td>0.000058</td><td>0.263158</td><td>-0.625535</td></tr>
<tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr>
</table>
</td>
</tr>
<tr>
<td><b>collections_12_mths_ex_med</b></td>
<td>This variable contains 98% of the data in one category, where both the failure and return rates are close to the overall average. The remaining 2% of the data is dispersed and does not provide conclusive insights. Therefore, we will remove this variable and thus prevent potential biases or inaccuracies in our analysis that could arise from including a variable with minimal variation and limited impact on the target variable.</td>
<td>

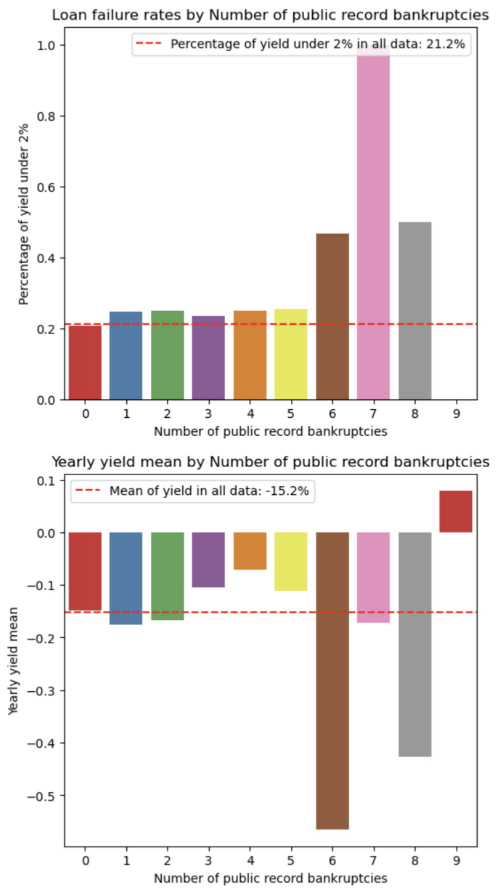


</td>
</tr>
</table>

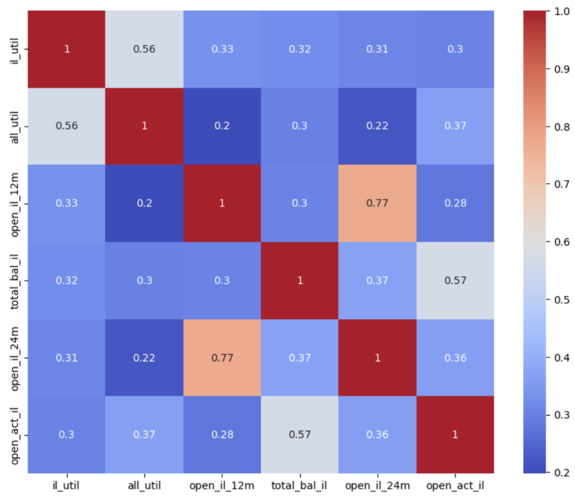
acc_now_delinq	This variable contains 99.36% of the data in one category, where both the failure and return rates are close to the overall average. Therefore, we will remove this variable and thus prevent potential biases or inaccuracies in our analysis that could arise from including a variable with minimal variation and limited impact on the target variable.	 Loan failure rates by Number of Currently Delinquent Accounts  Yearly yield mean by Number of Currently Delinquent Accounts
chargeoff_within_12_mths	This variable contains 99.17% of the data in one category, where both the failure and return rates are close to the overall average. Therefore, we will remove this variable and thus prevent potential biases or inaccuracies in our analysis that could arise from including a variable with minimal variation and limited impact on the target variable.	 Loan failure rates by Number of Charge-Offs in the Past 12 Months  Yearly yield mean by Number of Charge-Offs in the Past 12 Months

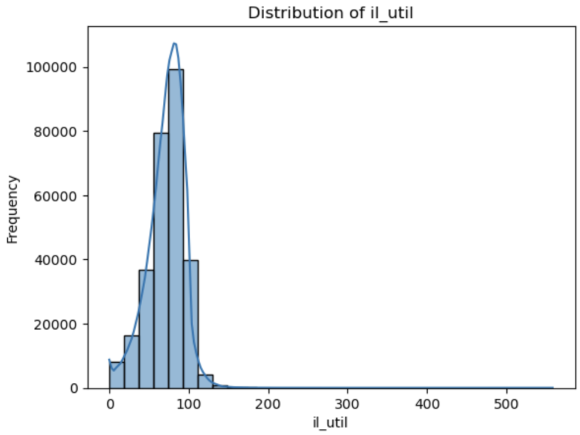
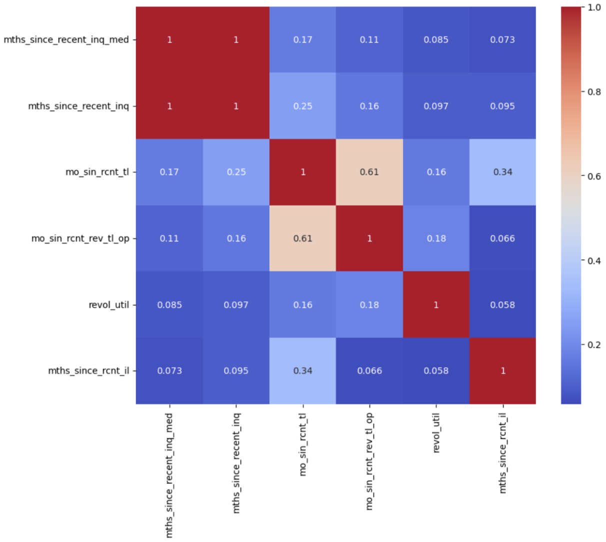
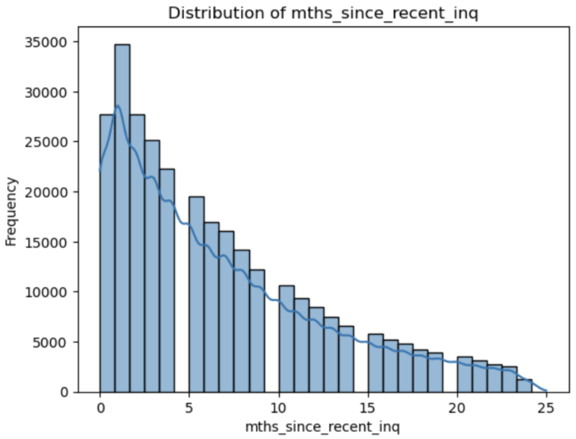
num_tl_120dpd_2m	This variable contains 99.89% of the data in one category, where both the failure and return rates are close to the overall average. Therefore, we will remove this variable and thus prevent potential biases or inaccuracies in our analysis that could arise from including a variable with minimal variation and limited impact on the target variable.	 
num_tl_30dpd	This variable contains 99.58% of the data in one category, where both the failure and return rates are close to the overall average. Therefore, we will remove this variable and thus prevent potential biases or inaccuracies in our analysis that could arise from including a variable with minimal variation and limited impact on the target variable.	 

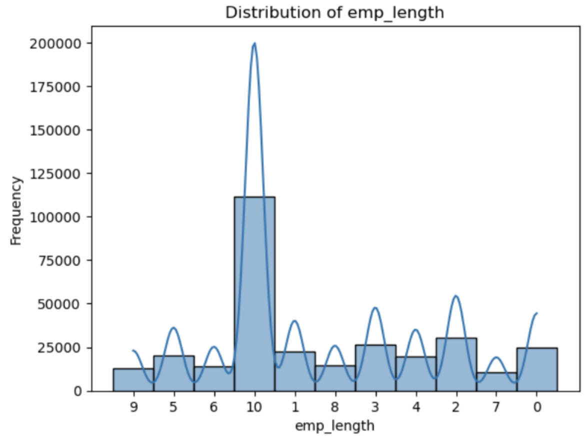
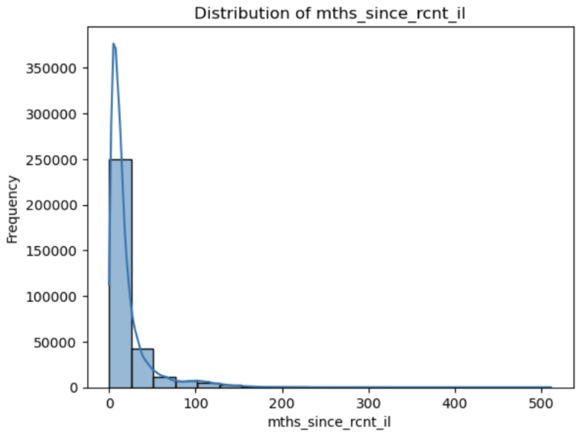
tax_liens	In this variable, 95% of the values are 0 tax liens that have similar rates of failure and average returns as the population, while higher values indicate a higher risk of default with lower average returns. This variable is relevant for predicting loan defaults because it indicates a borrower's history of tax liens, which may suggest difficulty in paying taxes and a higher risk of default. We converted to a binary variable where 0 represents no tax liens and 1 represents one or more tax liens.	 
num_tl_90g_dpd_24m	This variable has low entropy but is relevant for predicting loan defaults because a high number of accounts past due may indicate a history of not repaying debts on time, which increases the risk of default. While high numbers of accounts past due generally lead to higher failure rates, for certain cases with a high number of accounts past due, they do not represent a large percentage of the data. Therefore, we can make this variable binary, where 0 represents no accounts past due and 1 represents a history of having accounts past due for 90 or more days in the last 24 months.	 

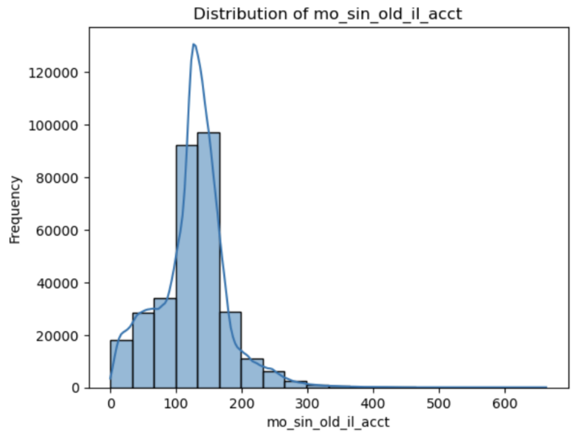
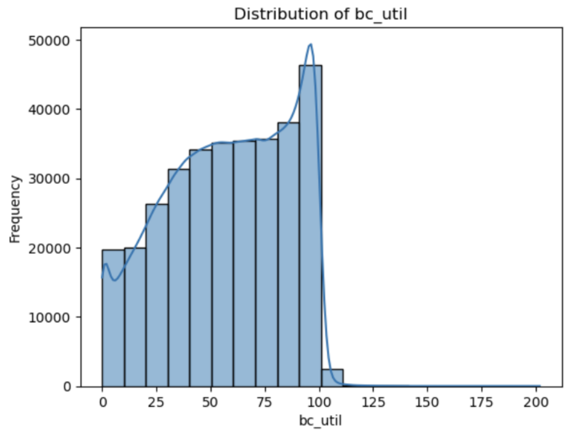
pub_rec_bankruptcies	This variable has low entropy but is relevant for predicting loan defaults because having a history of bankruptcy may indicate that the borrower has experienced financial difficulties in the past and may be at a higher risk of defaulting on the loan. Therefore, higher values of "pub_rec_bankruptcies" may be associated with a higher risk of default. Therefore, we can make this variable binary, where 0 represents no public record bankruptcies and 1 represents a more then one public record bankruptcies	

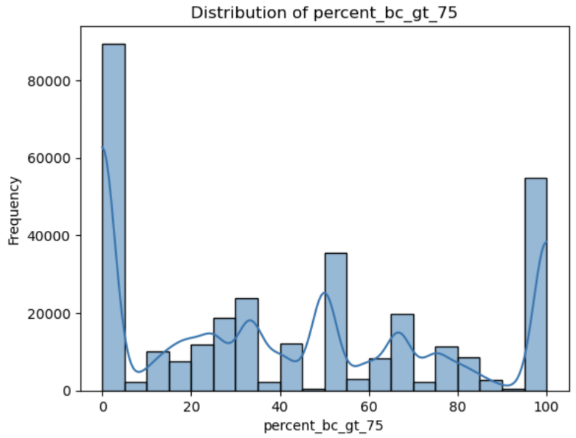
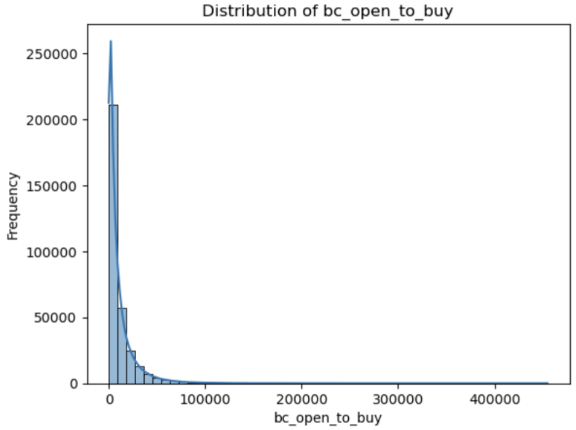
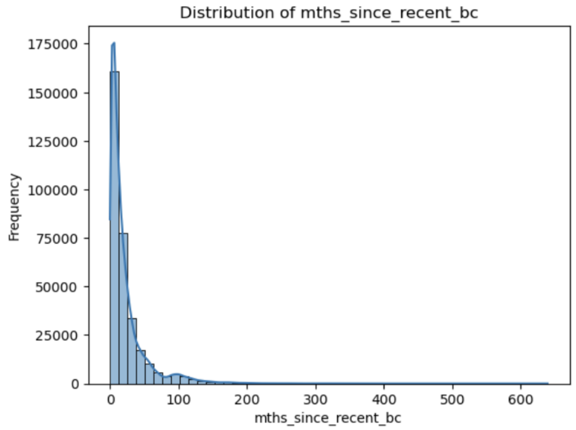
Appendix B:

Column Name	**Missing Values**	**% Missing**	**Imputation Method**	
il_util	43357	13.2%	LinearRegression with ['all_util', 'open_il_12m', 'total_bal_il', 'open_il_24m', 'open_act_il'] The "il_util" feature reflects a borrower's account utilization, which is important in assessing ability to return a loan. We can use other related columns such as payment and credit utilization to predict this variable, we can see that they have high correlation.	

			We observed that the distribution of "il_util" has a median greater than the mean, indicating a non-normal distribution. Hence, we will use the median to fill in the missing values, which is a more conservative approach. A lower value of "il_util" suggests that the borrower is utilizing a smaller portion of their available credit, potentially reducing the risk of default and improving their creditworthiness.	
mths_since_recent_inq	32274	9.82%	LinearRegression with ['mo_sin_rcnt_tl', 'mo_sin_rcnt_rev_tl_op', 'fico_range_high', 'fico_range_low'] The feature reflects the number of months since the borrower's most recent installment accounts opened, which is important in assessing ability to return a loan. We can use other related columns , we can see that there is high correlation.	
			The "mths_since_recent_inq" distribution shows a higher mean than the median, indicating some outliers in the data. To address missing values, we will use the median as a conservative approach. Higher values of "mths_since_rcnt_il" suggest that the borrower has not recently applied for new credit, which could reduce the risk of default.	

emp_length	21394	6.51%	We divided the annual_inc to bins and then generated all possible combinations of home ownership and income ranges, and computed the mode value of the available data for each combination.	
			As you can see, there is no clear distribution for the variable, but it seems that most of the values are 10, so we will use mode	
mths_since_rcn	8843	2.69%	The mean is larger than the median. This indicates that there are a few very large values (outliers) that are pulling up the mean. To be more conservative in our approach, we will use the average to impute missing values.	
			However, because the difference between the average and the median is relatively large due to the presence of outliers, we will also consider using the median to impute missing values to have a more robust and conservative approach.	

mo_sin_old_il_acct	8792	2.67%	The median is larger than the mean. This indicates that there are a few very large values (outliers) that are pulling up the mean. To be more conservative in our approach, we will use the median to impute missing values. However, because the difference between the average and the median is relatively large due to the presence of outliers, we will also consider using the mean to impute missing values to have a more robust and conservative approach.	
bc_util	3707	1.12%	The mean is larger than the median. This indicates that there are a few very large values (outliers) that are pulling up the mean. To be more conservative in our approach, we will use the average to impute missing values. However, because the difference between the average and the median is relatively large due to the presence of outliers, we will also consider using the median to impute missing values to have a more robust and conservative approach.	

percent_bc_gt_75	3567	1.08%	Because of the unclear distribution, we will complete both by mean and median.	
bc_open_to_buy	3542	1.07%	Because the mean and the median are very far from each other due to many outliers we will use both the mean and the median.	
mths_since_recent_bc	3328	1.01%	Because the mean and the median are very far from each other due to many outliers we will use both the mean and the median.	

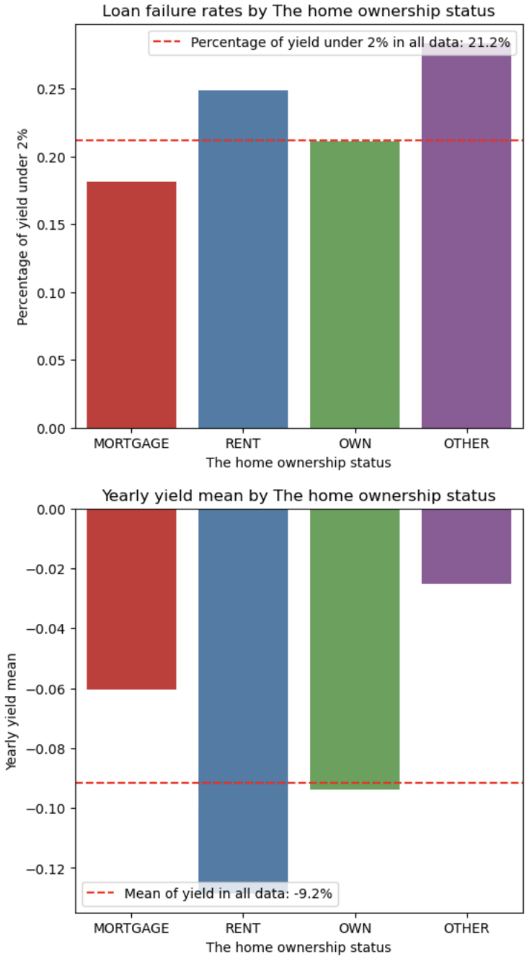
Appendix C:

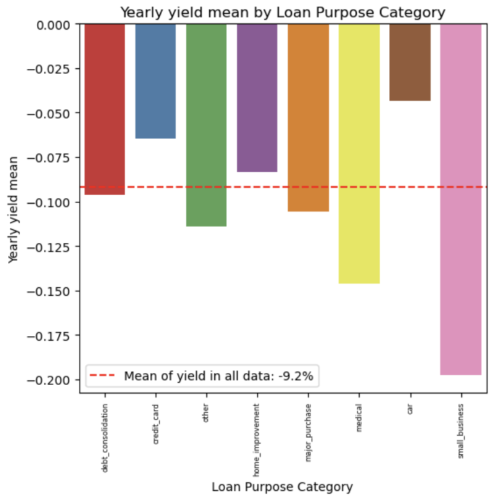
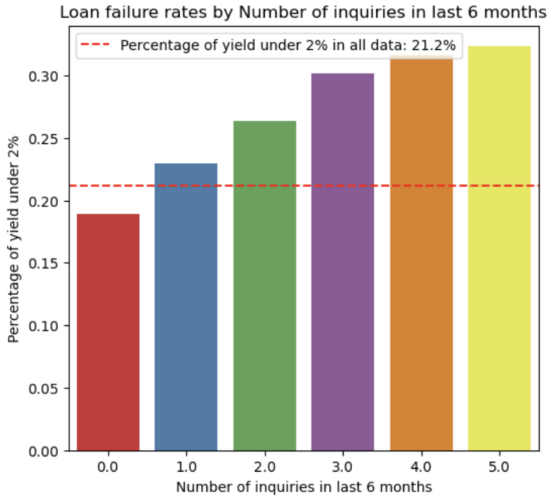
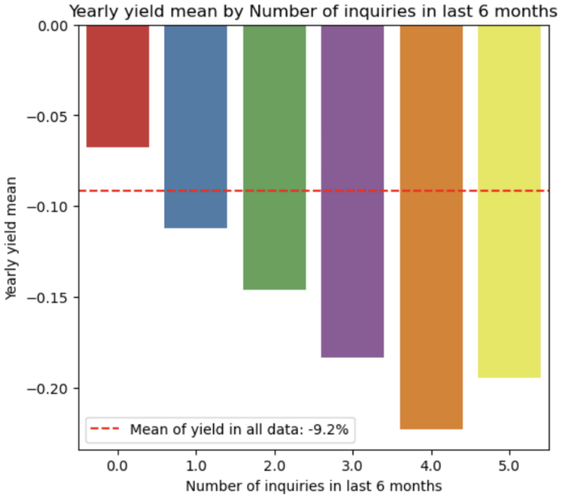
Column Name	**Transformation Details**		**New Column Name**
num_tl_90g_dpd_24m	Binary(0- no accounts, 1- a history of having accounts)	detailed in Appendix A	has_past_due_accounts_90g_24m
pub_rec_bankruptcies	Binary(0- no public record bankruptcies, 1- more than one public record bankruptcies)	detailed in Appendix A	has_pub_rec_bankruptcies
purpose	Binary(1- yield over 2% percentage greater than population mean,0- otherwise)	detailed in the step B report	purpose_danger
addr_state	Binary(1- yield over 2% percentage greater than population mean,0- otherwise)	detailed in the step B report	addr_state_danger
earliest_cr_line	data['issue_d'] - data["earliest_cr_line"]	detailed in the step B report	Age_of_credit_history
inq_last_6mths	Ordinal encoding (high-3,medium-2,low-1	detailed in Appendix D	inq_last_6mths_cat
inq_last_12m	data['inq_last_12m']-data['inq_last_6mths']	We want to capture the change in credit-seeking behavior of the borrower. If the difference is positive, it suggests an increase in credit-seeking behavior, which may be indicative of financial instability. Conversely, a negative difference could suggest a decrease in credit-seeking behavior, which may be indicative of greater financial stability. This new feature could potentially improve the model's ability to predict loan default risk.	recent_inquiry_difference

Mo_sin_rcnt_rev_tl_op and mo_sin_rcnt_tl	data['mo_sin_rcnt_rev_tl_op']/data['mo_sin_rcnt_tl']	We discovered that the correlation between the columns "mo_sin_rcnt_rev_tl_op" and "mo_sin_rcnt_tl" is 0.61- Thus, we created a new feature called "revolving_account_opened_ratio" by dividing "mo_sin_rcnt_rev_tl_op" by "mo_sin_rcnt_tl", aiming to capture the ratio of months since the most recent revolving account opened to months since the most recent account opened. The resulting feature may provide additional information about a borrower's credit behavior, which can be useful in predicting their ability to repay a loan.	revolving_account_opened_ratio

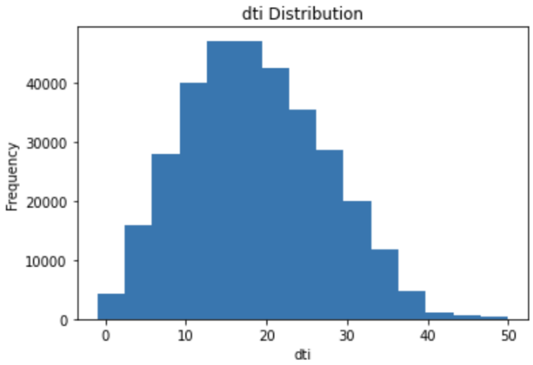
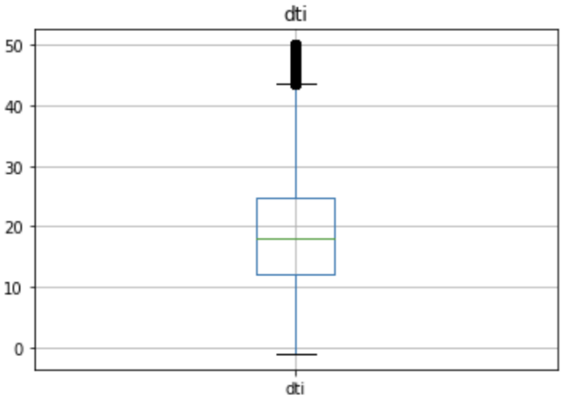
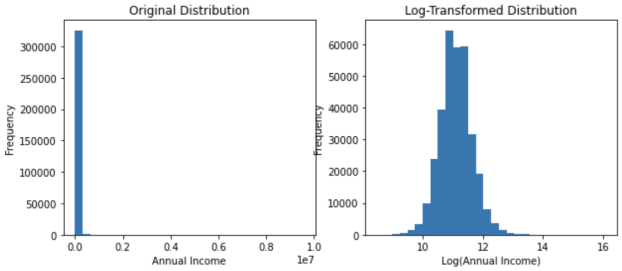
Appendix D:

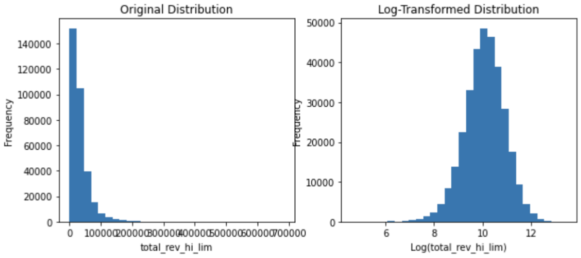
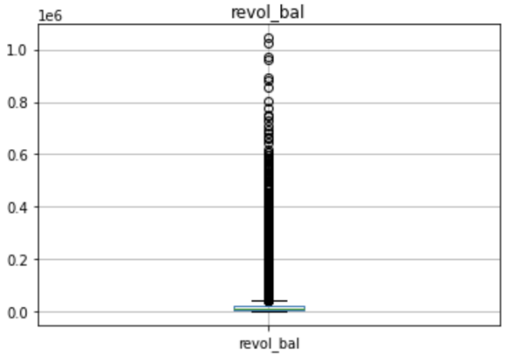
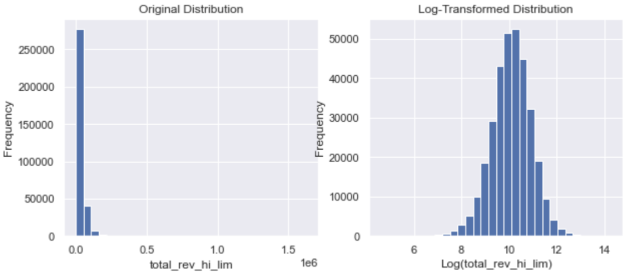
Note that many categorical variables already undergone transformations so reducing the number of variables requires further processing. You can refer to Appendix C.

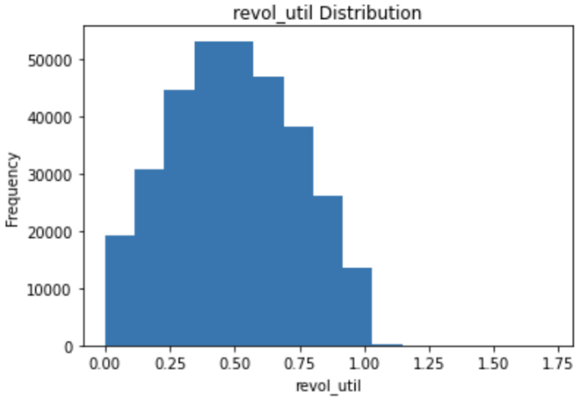
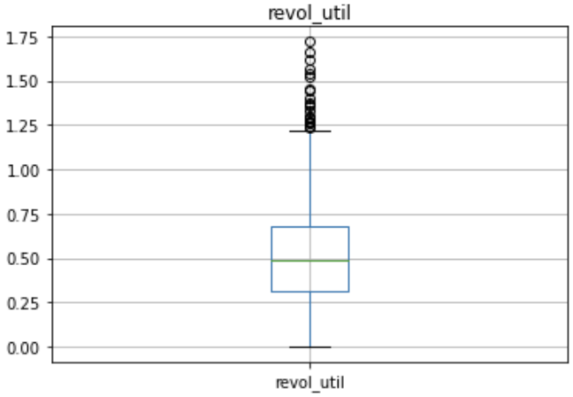
Column Name	**Imputation Method**	**Suitable Model**	
home_ownership	We will one-hot encode the categories RENT and MORTGAGE since they have a loan percentage above 2% and the average yield is above/below the mean in the data.	For both	

purpose	Leave One Out Encoding was used for the "purpose" variable in both regression and classification models due to its high cardinality and multiple categories. It helps prevent overfitting and improves model performance by using a more robust estimate of the mean for each category, making the models more generalizable.	For both	
inq_last_6mths	We are transforming this variable into an ordinal categorical variable with three levels based on the number of credit inquiries made in the past 6 months. This is done to capture the relationship between inquiry levels and default rates, which can be used as a predictor in a machine learning model to predict loan defaults. Ordinal encoding is appropriate because the categories have a natural order and relative values.	For both	 

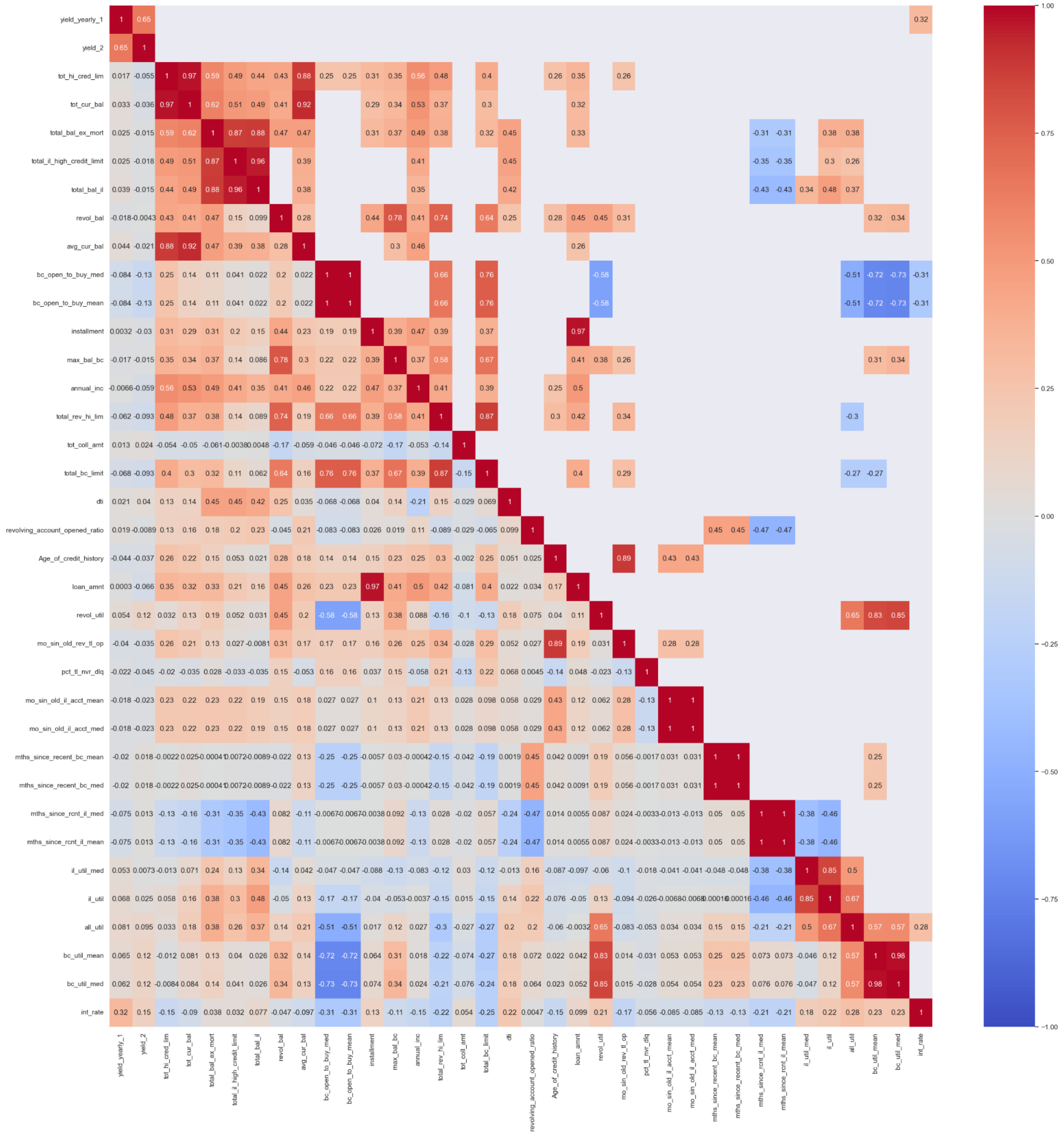
Appendix E:

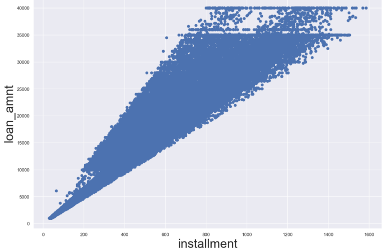
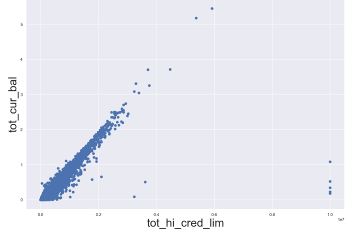
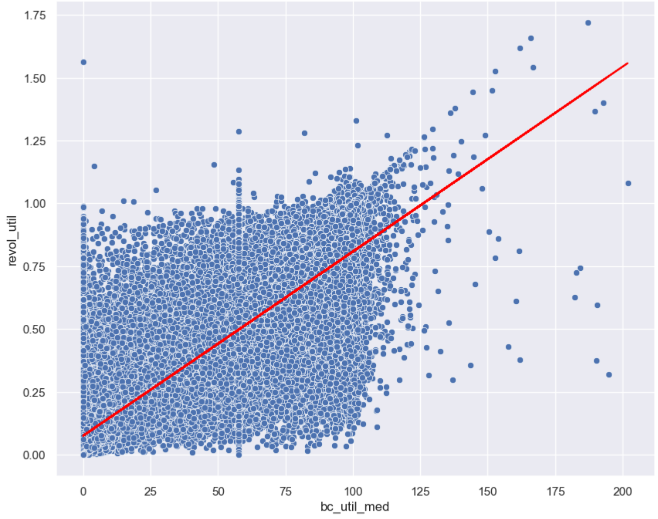
Column Name	**Treatment**	
DTI	After observing that the distribution of our data appears to be approximately normal, we used a boxplot to identify any potential outliers. By examining the boxplot, we looked for data points that fell outside the whiskers, which are typically set at a distance of 1.5 times the interquartile range (IQR). However, in our case, we decided to be more conservative and considered data points that were more than 3 standard deviations away from the mean as outliers.	
DTI		
annual_inc	when the distribution of a variable appears to be exponential or highly skewed, applying a logarithmic transformation can sometimes help to make the data more normally distributed. Taking the logarithm of the variable can compress the range of values,	

	reduce the skewness, and make the distribution more symmetric.	
Total rev limit	when the distribution of a variable appears to be exponential or highly skewed, applying a logarithmic transformation can sometimes help to make the data more normally distributed. Taking the logarithm of the variable can compress the range of values, reduce the skewness, and make the distribution more symmetric.	
revol_bal		
Revol_bal- num of outliers over 2000000	916 + when the distribution of a variable appears to be exponential or highly skewed, applying a logarithmic transformation can sometimes help to make the data more normally distributed. Taking the logarithm of the variable can compress the range of values, reduce the skewness, and make the distribution more symmetric.	

revol_util	After observing that the distribution of our data appears to be approximately normal, we used a boxplot to identify any potential outliers. By examining the boxplot, we looked for data points that fell outside the whiskers, which are typically set at a distance of 1.5 times the interquartile range (IQR). However, in our case, we decided to be more conservative and considered data points that were more than 3 standard deviations away from the mean as outliers.	 

Appendix F:


yield_yearly_1 1 0.65 0.32
yield_2 0.65 1
tot_hi_cred_lim 0.017 -0.055 1 0.97 0.59 0.49 0.44 0.43 0.88 0.25 0.25 0.31 0.35 0.56 0.48 0.4 0.26 0.35 0.26
tot_cur_bal 0.033 -0.036 0.97 1 0.62 0.51 0.49 0.41 0.92 0.29 0.34 0.53 0.37 0.3 0.32
total_bal_ex_mort 0.025 -0.015 0.59 0.62 1 0.87 0.88 0.47 0.47 0.31 0.37 0.49 0.38 0.32 0.45 0.33 -0.31 -0.31 0.38 0.38
total_il_high_credit_limit 0.025 -0.018 0.49 0.51 0.87 1 0.96 0.39 0.41 0.45 -0.35 -0.35 0.3 0.26
total_bal_il 0.039 -0.015 0.44 0.49 0.88 0.96 1 0.38 0.35 0.42 -0.43 -0.43 0.34 0.48 0.37
revol_bal -0.018 -0.0043 0.43 0.41 0.47 0.15 0.099 1 0.28 0.44 0.78 0.41 0.74 0.64 0.25 0.28 0.45 0.45 0.31 0.32 0.34
avg_cur_bal 0.044 -0.021 0.88 0.92 0.47 0.39 0.38 0.28 1 0.3 0.46 0.26
bc_open_to_buy_med -0.084 -0.13 0.25 0.14 0.11 0.041 0.022 0.2 0.022 1 1 0.66 0.76 -0.58 -0.51 -0.72 -0.73 -0.31
bc_open_to_buy_mean -0.084 -0.13 0.25 0.14 0.11 0.041 0.022 0.2 0.022 1 1 0.66 0.76 -0.58 -0.51 -0.72 -0.73 -0.31
installment 0.0032 -0.03 0.31 0.29 0.31 0.2 0.15 0.44 0.23 0.19 0.19 1 0.39 0.47 0.39 0.37 0.97
max_bal_bc -0.017 -0.015 0.35 0.34 0.37 0.14 0.086 0.78 0.3 0.22 0.22 0.39 1 0.37 0.58 0.67 0.41 0.38 0.26 0.31 0.34
annual_inc -0.0066 -0.059 0.56 0.53 0.49 0.41 0.35 0.41 0.46 0.22 0.22 0.47 0.37 1 0.41 0.39 0.25 0.5
total_rev_hi_lim -0.062 -0.093 0.48 0.37 0.38 0.14 0.089 0.74 0.19 0.66 0.66 0.39 0.58 0.41 1 0.87 0.3 0.42 0.34 -0.3
tot_coll_amt 0.013 0.024 -0.054 -0.05 -0.061 -0.0038 0.0048 -0.17 -0.059 -0.046 -0.046 -0.072 -0.17 -0.053 -0.14 1
total_bc_limit -0.068 -0.093 0.4 0.3 0.32 0.11 0.062 0.64 0.16 0.76 0.76 0.37 0.67 0.39 0.87 -0.15 1 0.4 0.29 -0.27 -0.27
dti 0.021 0.04 0.13 0.14 0.45 0.45 0.42 0.25 0.035 -0.068 -0.068 0.04 0.14 -0.21 0.15 -0.029 0.069 1
revolving_account_opened_ratio 0.019 -0.0089 0.13 0.16 0.18 0.2 0.23 -0.045 0.21 -0.083 -0.083 0.026 0.019 0.11 -0.089 -0.029 -0.065 0.099 1 0.45 0.45 -0.47 -0.47
Age_of_credit_history -0.044 -0.037 0.26 0.22 0.15 0.053 0.021 0.28 0.18 0.14 0.14 0.15 0.23 0.25 0.3 -0.002 0.25 0.051 0.025 1 0.89 0.43 0.43
loan_amnt 0.0003 -0.066 0.35 0.32 0.33 0.21 0.16 0.45 0.26 0.23 0.23 0.97 0.41 0.5 0.42 -0.081 0.4 0.022 0.034 0.17 1
revol_util 0.054 0.12 0.032 0.13 0.19 0.052 0.031 0.45 0.2 -0.58 -0.58 0.13 0.38 0.088 -0.16 -0.1 -0.13 0.18 0.075 0.04 0.11 1 0.65 0.83 0.85
mo_sin_old_rev_tl_op -0.04 -0.035 0.26 0.21 0.13 0.027 -0.0081 0.31 0.17 0.17 0.17 0.16 0.26 0.25 0.34 -0.028 0.29 0.052 0.027 0.89 0.19 0.031 1 0.28 0.28
pct_tl_nvr_dlq -0.022 -0.045 -0.02 -0.035 0.028 -0.033 -0.035 0.15 -0.053 0.16 0.16 0.037 0.15 -0.058 0.21 -0.13 0.22 0.068 0.0045 -0.14 0.048 -0.023 -0.13 1
mo_sin_old_il_acct_mean -0.018 -0.023 0.23 0.22 0.23 0.22 0.19 0.15 0.18 0.027 0.027 0.1 0.13 0.21 0.13 0.028 0.098 0.058 0.029 0.43 0.12 0.062 0.28 -0.13 1 1
mo_sin_old_il_acct_med -0.018 -0.023 0.23 0.22 0.23 0.22 0.19 0.15 0.18 0.027 0.027 0.1 0.13 0.21 0.13 0.028 0.098 0.058 0.029 0.43 0.12 0.062 0.28 -0.13 1 1
mths_since_recent_bc_mean -0.02 0.018 -0.0022 0.025 -0.00041 0.0072 -0.0089 -0.022 0.13 -0.25 -0.25 -0.0057 0.03 -0.00042 -0.15 -0.042 -0.19 0.0019 0.45 0.042 0.0091 0.19 0.056 -0.0017 0.031 0.031 1 1 0.25
mths_since_recent_bc_med -0.02 0.018 -0.0022 0.025 -0.00041 0.0072 -0.0089 -0.022 0.13 -0.25 -0.25 -0.0057 0.03 -0.00042 -0.15 -0.042 -0.19 0.0019 0.45 0.042 0.0091 0.19 0.056 -0.0017 0.031 0.031 1 1 0.25
mths_since_rcnt_il_med -0.075 0.013 -0.13 -0.16 -0.31 -0.35 -0.43 0.082 -0.11 -0.0067 -0.0067 -0.0038 0.092 -0.13 0.028 -0.02 0.057 -0.24 -0.47 0.014 0.0055 0.087 0.024 -0.0033 -0.013 -0.013 0.05 0.05 1 1 -0.38 -0.46
mths_since_rcnt_il_mean -0.075 0.013 -0.13 -0.16 -0.31 -0.35 -0.43 0.082 -0.11 -0.0067 -0.0067 -0.0038 0.092 -0.13 0.028 -0.02 0.057 -0.24 -0.47 0.014 0.0055 0.087 0.024 -0.0033 -0.013 -0.013 0.05 0.05 1 1 -0.38 -0.46
il_util_med 0.053 0.0073 -0.013 0.071 0.24 0.13 0.34 -0.14 0.042 -0.047 -0.047 -0.088 -0.13 -0.083 -0.12 0.03 -0.12 -0.013 0.16 -0.087 -0.097 -0.06 -0.1 -0.018 -0.041 -0.041 -0.048 -0.048 -0.38 -0.38 1 0.85 0.5
il_util 0.068 0.025 0.058 0.16 0.38 0.3 0.48 -0.05 0.13 -0.17 -0.17 -0.04 -0.053 -0.0037 -0.15 0.015 -0.15 0.14 0.22 -0.076 -0.05 0.13 -0.094 -0.026 -0.0068 -0.0068 0.00016 0.00016 -0.46 -0.46 0.85 1 0.67
all_util 0.081 0.095 0.033 0.18 0.38 0.26 0.37 0.14 0.21 -0.51 -0.51 0.017 0.12 0.027 -0.3 -0.027 -0.27 0.2 0.2 -0.06 -0.0032 0.65 -0.083 -0.053 0.034 0.034 0.15 0.15 -0.21 -0.21 0.5 0.67 1 0.57 0.57 0.28
bc_util_mean 0.065 0.12 -0.012 0.081 0.13 0.04 0.026 0.32 0.14 -0.72 -0.72 0.064 0.31 0.018 -0.22 -0.074 -0.27 0.18 0.072 0.022 0.042 0.83 0.014 -0.031 0.053 0.053 0.25 0.25 0.073 0.073 -0.046 0.12 0.57 1 0.98
bc_util_med 0.062 0.12 -0.0084 0.084 0.14 0.041 0.026 0.34 0.13 -0.73 -0.73 0.074 0.34 0.024 -0.21 -0.076 -0.24 0.18 0.064 0.023 0.052 0.85 0.015 -0.028 0.054 0.054 0.23 0.23 0.076 0.076 -0.047 0.12 0.57 0.98 1
int_rate 0.32 0.15 -0.15 -0.09 0.038 0.032 0.077 -0.047 -0.097 -0.31 -0.31 0.13 -0.11 -0.15 -0.22 0.054 -0.25 0.22 0.0047 -0.15 0.099 0.21 -0.17 -0.056 -0.085 -0.085 -0.13 -0.13 -0.21 -0.21 0.18 0.22 0.28 0.23 0.23 1
yield_yearly_1
yield_2
tot_hi_cred_lim
tot_cur_bal
total_bal_ex_mort
total_il_high_credit_limit
total_bal_il
revol_bal
avg_cur_bal
bc_open_to_buy_med
bc_open_to_buy_mean
installment
max_bal_bc
annual_inc
total_rev_hi_lim
tot_coll_amt
total_bc_limit
dti
revolving_account_opened_ratio
Age_of_credit_history
loan_amnt
revol_util
mo_sin_old_rev_tl_op
pct_tl_nvr_dlq
mo_sin_old_il_acct_mean
mo_sin_old_il_acct_med
mths_since_recent_bc_mean
mths_since_recent_bc_med
mths_since_rcnt_il_med
mths_since_rcnt_il_mean
il_util_med
il_util
all_util
bc_util_mean
bc_util_med
int_rate
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Column Name	**Treatment**	
installment	To reduce multicollinearity due to a very strong correlation (0.97, Spearman correlation) between ‘installment’ and ‘loan_amnt’ we decided to drop installment.	Scatter Plot: Relationship between installment and loan_amnt: strong positive association 
tot_hi_cred_lim	To reduce the number of features and also reduce multicollinearity due to a very strong correlation (0.92, Spearman correlation) between ‘tot_cur_bal’ and ‘tot_hi_cred_lim’ we decided to drop installment.	Scatter Plot: Relationship between "tot_hi_cred_lim" and "tot_cur_bal": strong positive association 
Bc_util_med - Ratio of total current balance to high credit/credit limit for all bankcard accounts. Revol_util - Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.	Both measures essentially reflect the borrower's credit utilization behavior, but at different levels of aggregation. The ratio of total current balance to high credit/credit limit focuses specifically on bankcard accounts, while the revolving line utilization rate considers all available revolving credit. We created a Credit_Utilization_Score feature that represents the combined credit utilization across bankcard accounts.	Scatter Plot: Relationship between "bc_util_med" and "revol_util" Polynomial Regression  revol_util 0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75 0 25 50 75 100 125 150 175 200 bc_util_med