INTEREST RATE PREDICTION

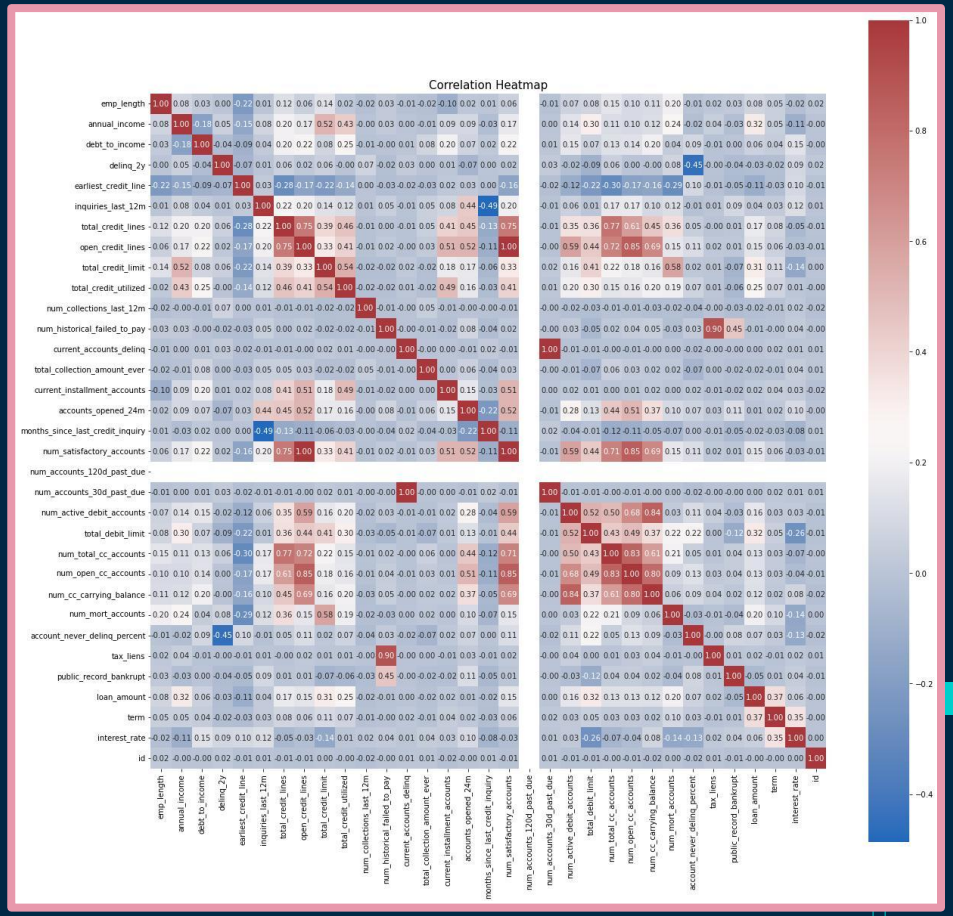
~ TEAM SCKNN ~

Sushil Deore | Nancy Luong | Natalia Mora | Kenneth Mccarthy | Charan Kanwal Preet Singh

Data Exploration Analysis

Understand data using the process of trial and errors including:

- Non-graphical analysis
- Univariate Analysis
- Multivariate Analysis



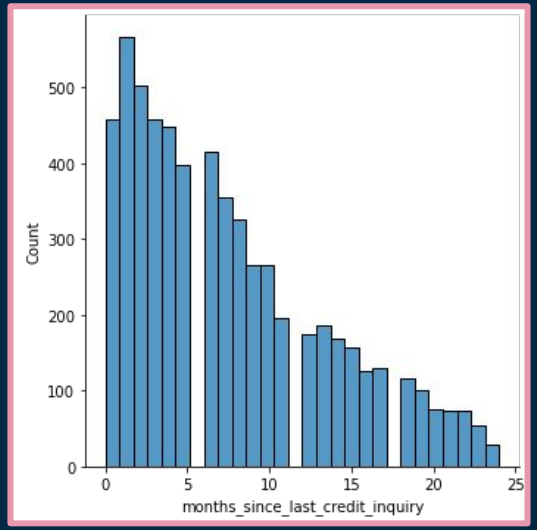
Data Cleaning & Preprocessing

Handling Null Values:

- Full drop of attribute if >50% NA
- Attribute distributions to determine proper imputation method

Qualitative Attributes:

- Outlier treatment
- Min/max scaling to normalize data
- Impute with median value due to skewed data



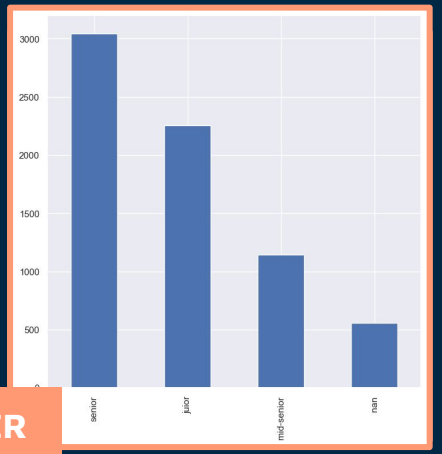
Quantitative Attributes:

- Recognizing/transforming discrete numeric attributes to categorical
- Condensing underrepresented values into "other" categories
- Example:

Attribute appear categorical, but deeper exploration reveals numerical data type

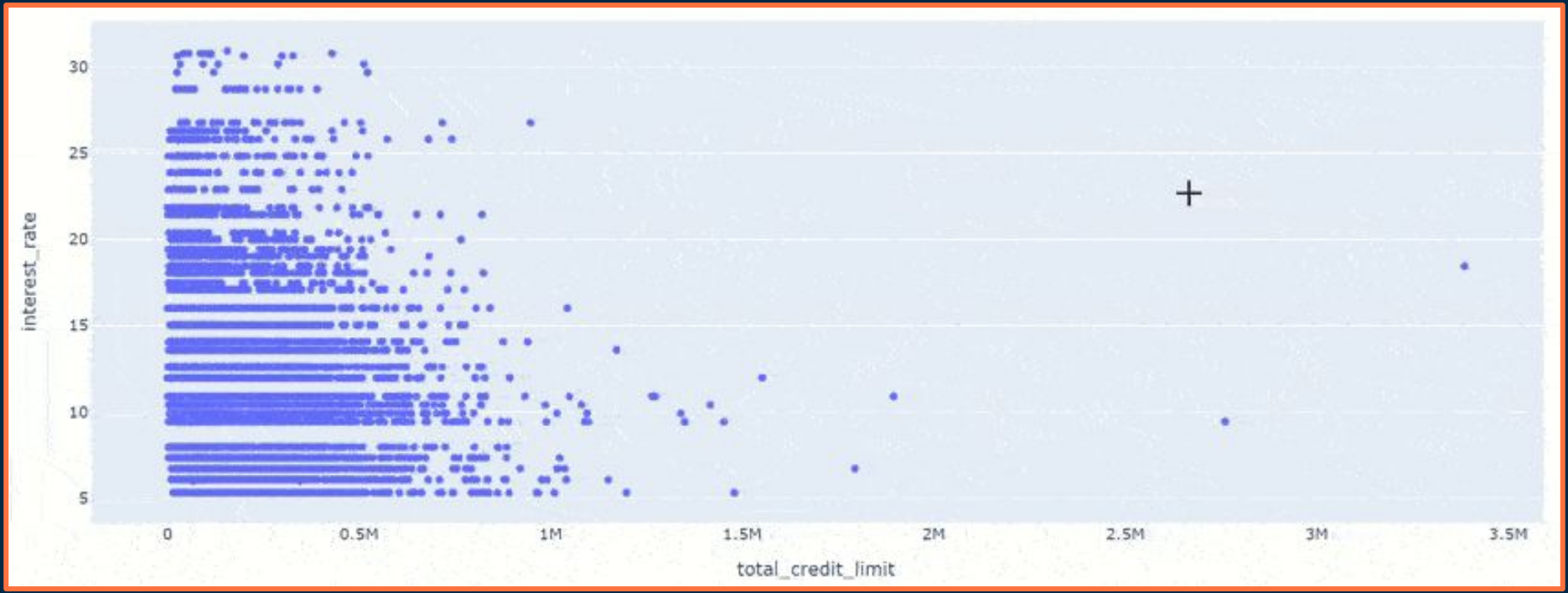
Imputing using MEDIAN significantly affects distribution

>> Best Option: binning column with additional sublevel for missing values



AFTER

Data Visualization



Raw Quantitative variables versus Target Variable (Interest Rate)

Feature Engineering

- Perform feature engineering In order to reduce dimensionality of model and improve accuracy
- Perform **one hot encoding** to be able to include categorical variable in model.
  - Convert categorical variables into several binary valued columns

	homeownership	OWN	RENT
0	RENT	0	1
1	RENT	0	1
2	OWN	1	0
3	MORTGAGE	0	0
4	RENT	0	1

- Identify columns that have high correlation with each other and **eliminate redundancy**

Modeling

Model: Several algorithms were created to determine the most accurate predictor. These models included:

- Linear Regression
  - WIth Recursive Feature Elimination
  - With LassoCV Feature Selection
  - With ElasticNetCV
- Decision Trees
- AdaBoost
- Gradient Boost

Ultimately Random Forest Regression model yielded the most accurate results

Model Justification: Random Forest produced the best R2 score of all models tested

Hyperparameter Tuning

The hyperparameters of the Random Forest Regression model are max_depth, n_jobs, and min_samples_leaf.

These were optimized using a randomized grid search to iterate over many combinations of hyperparameters in order to select the ones that led to the best model

Accuracy Report

Metric: The R2 score was used as a metric for judging the model

- Model had R2 of 0.39 for training data
- Model had R2 of 0.34 for test data

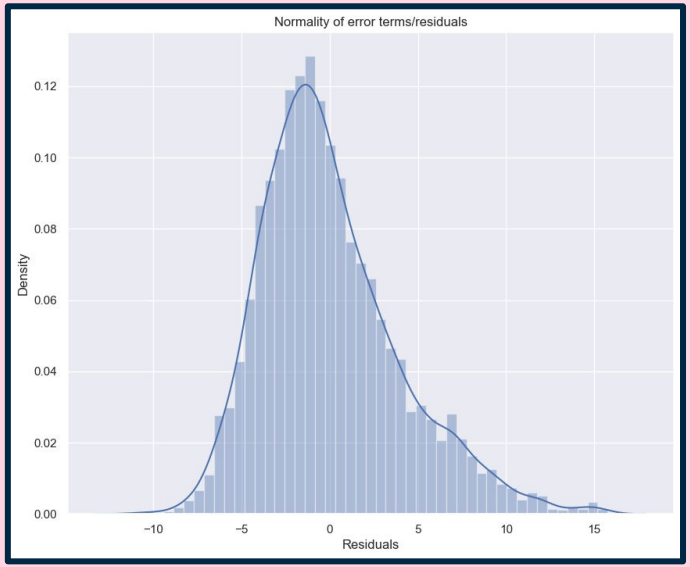
Metric Justification:

The R2 metric provides a simple to judge metric between 0 and 1 to easily judge model quality.

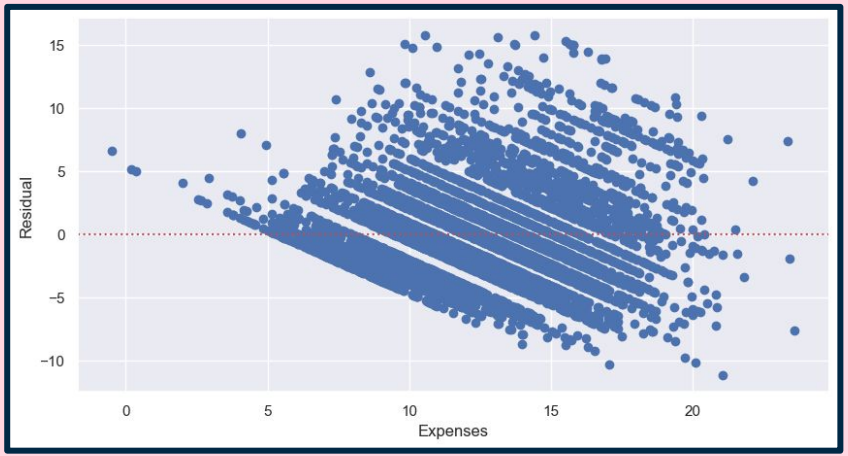
Accuracy Report

Residual Analyses

Distribution of error terms on training dataset:



Residuals Versus Predictions:



Conclusion

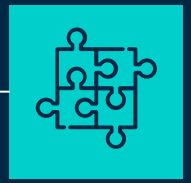
Random Forest Regression model gives the best prediction for this dataset

The table to the right displays the top 10 attributes that hold the most weight in predicting interest rate.

- Term
- Total Debt Limit: this correlation makes intuitive sense as limits are determined based on an individual's financial history and therefore is likely to reflect their likelihood to show financial responsibility

	Varname	Imp
16	term	0.335442
10	total_debit_limit	0.267897
1	debt_to_income	0.134740
38	whole	0.083813
23	Verified	0.035654
11	num_mort_accounts	0.032809
8	accounts_opened_24m	0.022491
12	account_never_delinq_percent	0.018841
15	loan_amount	0.011578
39	DirectPay	0.011453

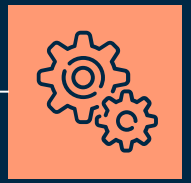
Conclusion



01

PROBLEM & SOLUTION

Here you could describe the topic of the section



02

OUR PROCESS

Here you could describe the topic of the section

03

TARGET

Here you could describe the topic of the section