Data Scientist 2025

January 2025

Data Overview

01

Datasets Summary

Application Data

- **647** total customers
- **30** Features
- Primary Key: Customer ID
- Earliest Loan: October 16, 2010
- Latest Loan: April 17, 2011

Loan Performance

- **1269** total loans with recorded performances ‘Good’ or ‘Bad’
- **634** customers match with at least 1 loan with performance history
- Primary Key: Loan ID = Customer ID-Loan # (e.g. a1b2345cde-02)

Zillow Home Value Index (External)

- Estimated home values in 50 states of the US and District of Columbia.
- Home Value recorded every **month**
- Earliest Home Value: Jan 2000
- Latest Home Value: Sep 2024

Excluded Predictors

High Missingness

other_phone_type

	Number of Missing Values	Proportion Missing (%)
payment_amount_approved	20	3.09%
bank_account_duration	1	0.15%
other_phone_type	285	44.05%
how_use_money	2	0.31%

Uninformative without more data

payment_ach, status

monthly_rent_amount

address_zip, email

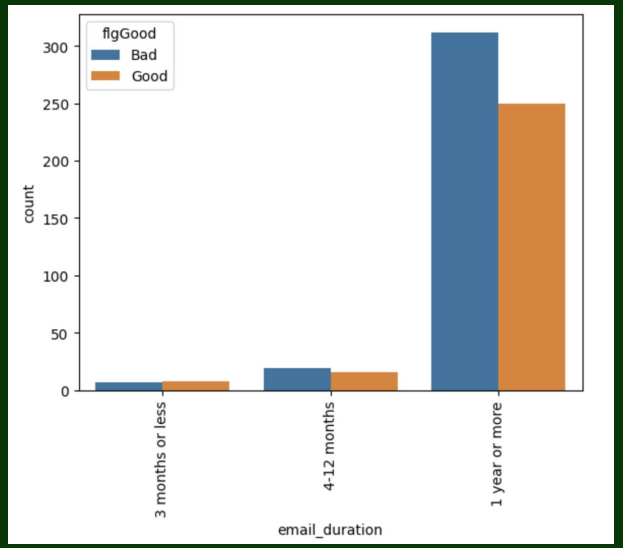
bank_routing_number

Included Predictors

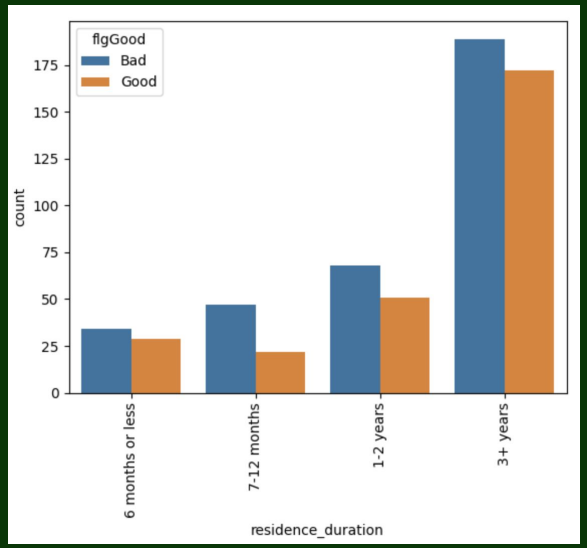
	Numerical	Categorical
Loan-Specific Features	• amount requested • amount approved • loan duration • duration approved	
Credit Score/Spending Features	• l2c score • FICO telecom • FICO retail • FICO bank card • FICO money	• how use money
Income Features	• age • monthly income amount	• residence duration • residence rent or own
Payment Features	• payment amount • monthly rent amount • payment amount approved • payment frequency • num_payments	• payment frequency • bank account direct deposit • bank account duration
Miscellaneous Features		• region • email duration • home phone type

Duration Visualizations

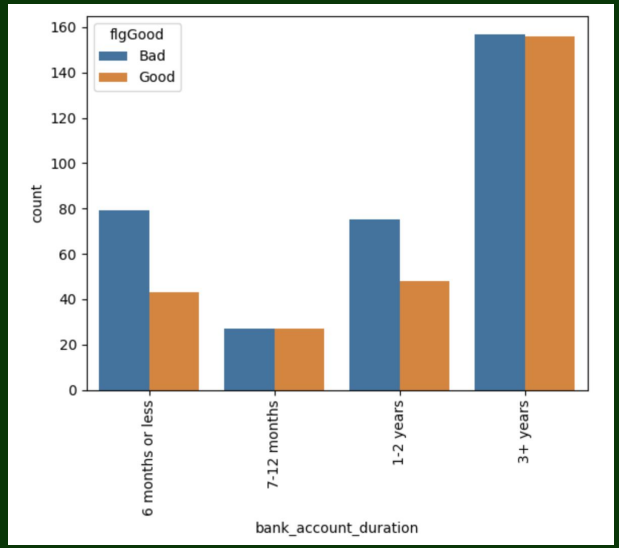
Email Duration



Residence Duration

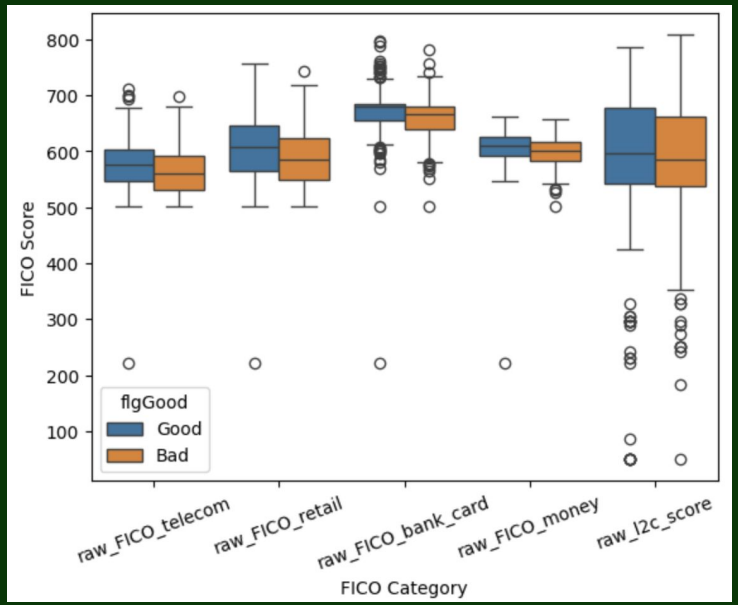


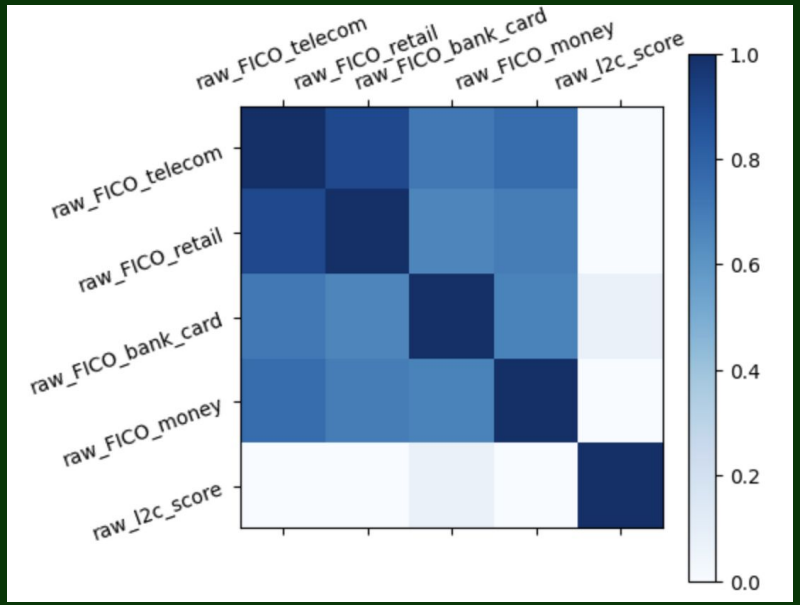
Bank Account Duration



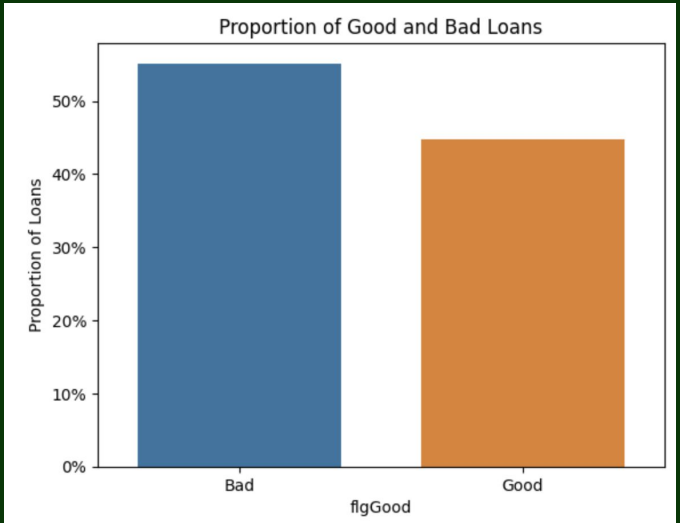
FICO Visualizations

FICO Scores





Loan Performance (Target Variable)



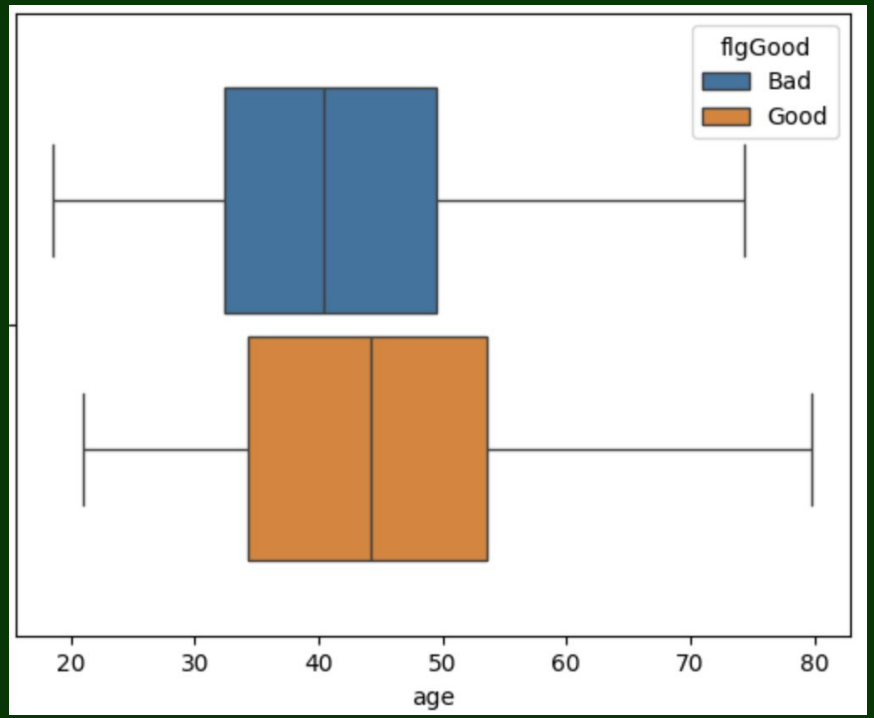
Feature Engineering and Selection

02

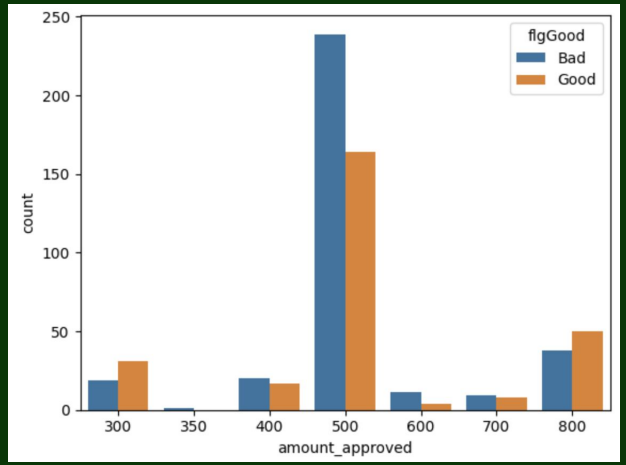
Features Engineered

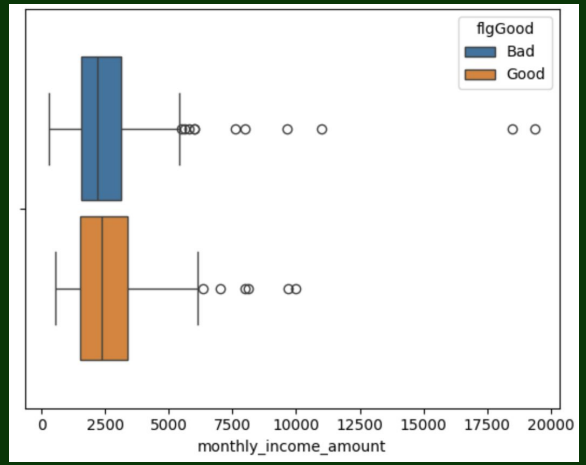
	Age	Debt to Income	Loan Approval Ratio	Payment Approval Ratio	Most Recent Loan	Zillow Home Value Index (ZHVI)	Customer Region
How it was derived	Datetime difference from birth date to time of application.	Divide the loan approved amount (monthly)+monthly rent amount by the monthly income amount.	Divide the loan approved amount by the loan requested amount.	Divide the loan payment approved amount by the loan payment amount.	From the loan performance dataframe, retrieve the most recent loan performance in the lending history of each customer.	Retrieved the most recent ZHVI prior to customer loan application date from external dataset.	Retrieved from ZIP to State to Region mapping.

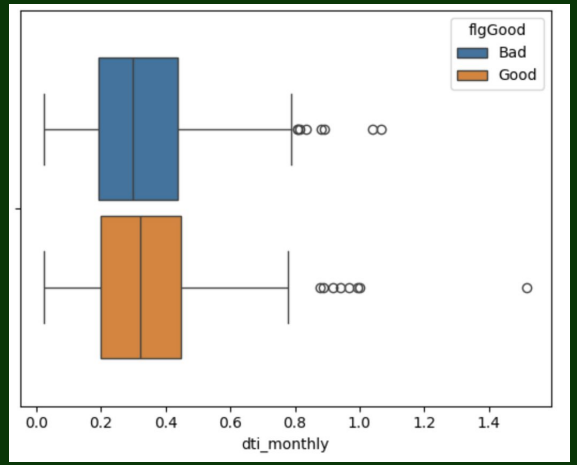
Age



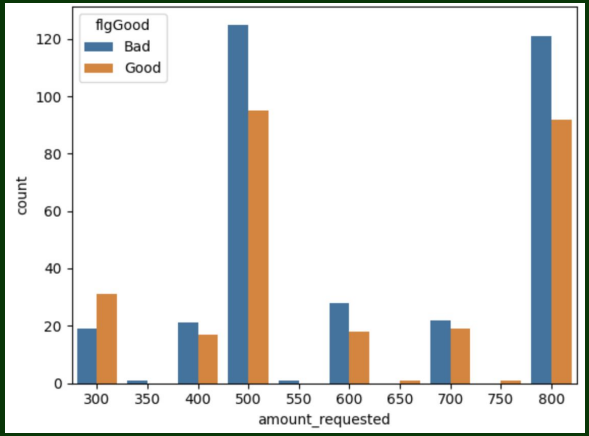
Loan Debt to Income

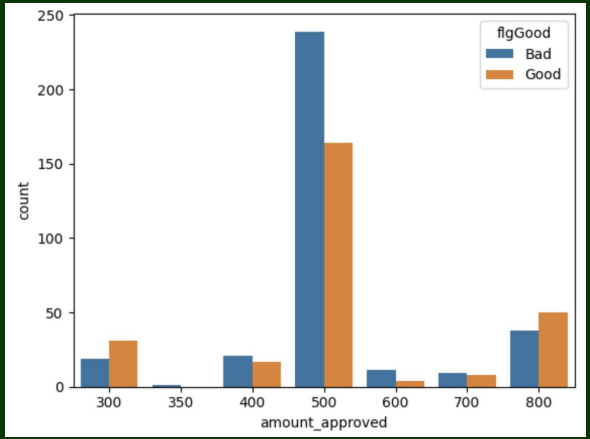


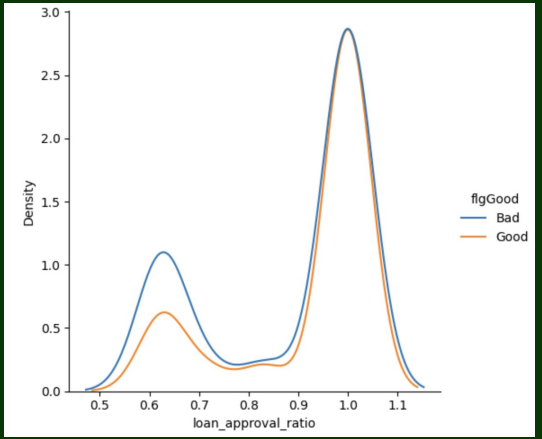




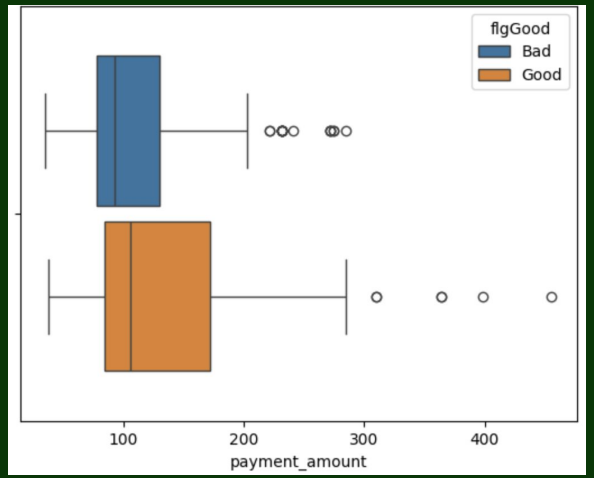
Loan Amount Approval

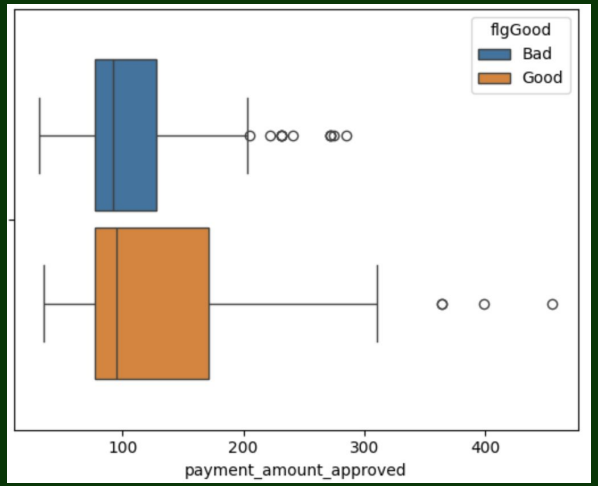


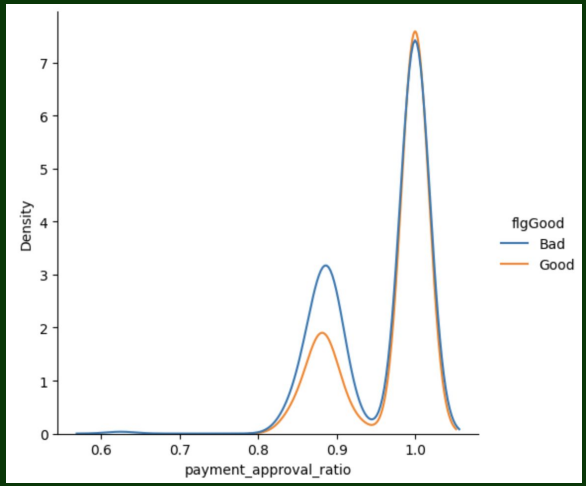




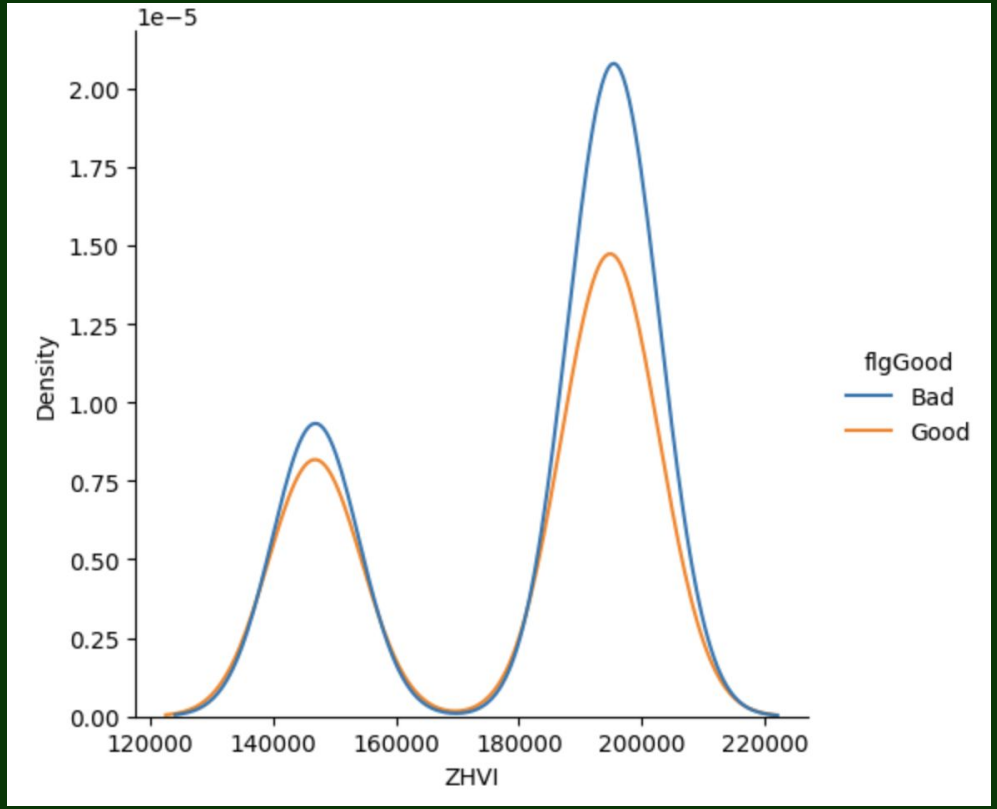
Loan Payment Approval



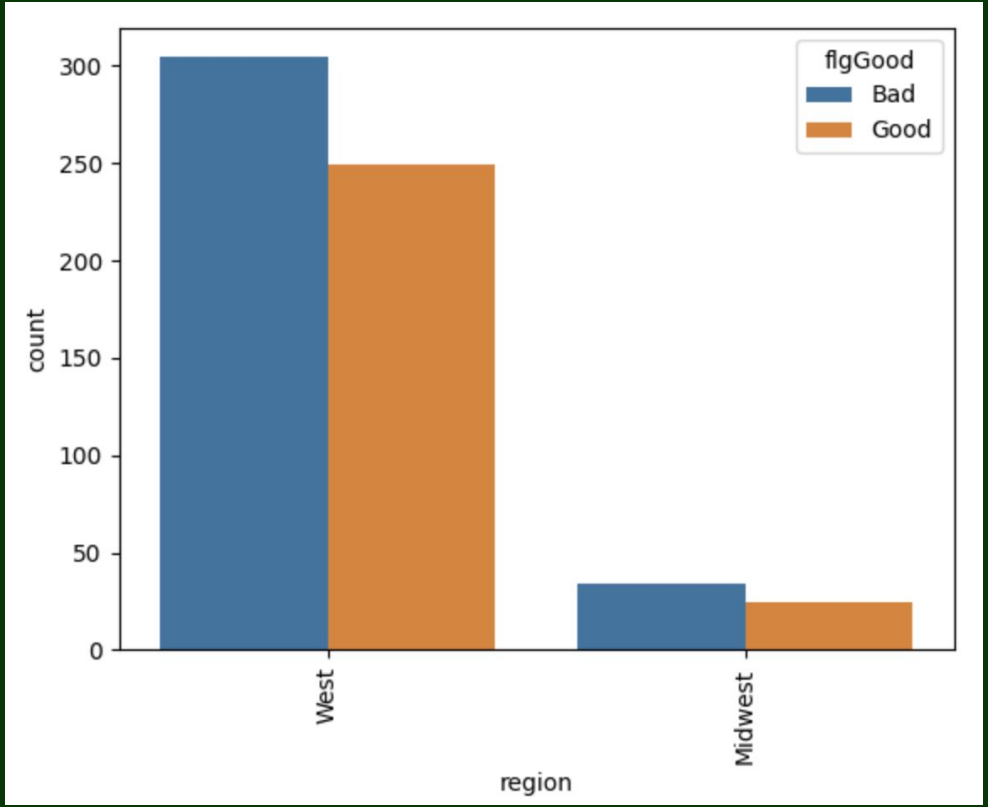




ZHVI Distribution of the loans



Customer Region



Features Selection Methods

5-fold Cross Validation	Lasso Feature Selection	Random Forest Features By Importance (From Top 20)	Forward Stepwise Feature Selection (SFS)
Features about loan	• amount requested • duration approved	• amount requested • amount approved • loan duration • duration approved	• loan duration • duration approved • amount approved
Credit Score Features	• FICO retail • FICO bank card	• FICO retail • FICO bank card • FICO money • FICO telecom • I2c score	• FICO retail • FICO bank card • FICO money • FICO telecom
Income Features	• ZHVI • age • residence duration • residence rent/own	• ZHVI • age • Debt to Income	• age • Debt to Income • residence duration
Payment Features	• bank account duration • payment frequency • payment approved	• bank account duration • payment amount approved • payment amount • num payments • payment frequency	• bank account duration • payment amount approved • payment amount • num payments • payment frequency
Other Features			• how use money, home phone type, region

Modeling

03

Models Considered

	Logistic Regression	Decision Tree	Random Forest	XGBoost
Reason for Consideration	Interpretable coefficients, baseline	Can capture non-linearity, interpretable, and resembles decision-making when deciding to lend - series of thresholds	Reduces the overfitting in a single decision tree and correlations between trees	Can accurately predict loan performance due to optimizing misclassifications of multiple weak learners

*Each of the 4 models are trained on 70% of 612 loans (428 loans) after excluding missingness and tested on 30% (184 loans).

Hyperparameter Search Space

Logistic Regression	Decision Tree	Random Forest	XGBoost
‘C’: [0.1, 0.5, 0.75, 1, 2, 3]	‘max_depth’: [None, 3, 5, 7, 9, 11]	‘n_estimators’: [10, 15, 20, 30, 40, 50]	’n_estimators’: [75, 100, 125]
”penalty”: [”elasticnet”]	’min_samples_split’: [2, 5, 10]	’max_depth’: [None, 3, 5, 7, 9, 11]	’eta’: [0.25, 0.4, 0.55]
’l1_ratio’: np.arange(0,1,0.05)	’min_samples_leaf’: [1, 2, 4]	’n_estimators’: [75, 100, 125]	’gamma’: [0.01, 0.1, 0.2]
’solver’: [’saga’]	’max_leaf_nodes’: [None, 10, 20]		’max_depth’: [None]
”max_iter”:[5000]			’lambda’: [0.01, 0.1, 0.2]
			’alpha’: [0.01, 0.1, 0.2]
			’subsample’: [0.35, 0.5, 0.65]

Selected Hyperparameters

Logistic Regression	Decision Tree	Random Forest	XGBoost
‘C’: 0.5	‘max_depth’: 11	‘n_estimators’: 30	’n_estimators’: 75
”penalty”: ”elasticnet”	’min_samples_split’: 2	’max_depth’: 9	’eta’: 0.25
’l1_ratio’: 0.75	’min_samples_leaf’: 4		’gamma’: 0.2
’solver’: ‘saga’			’lambda’: 0.01
”max_iter”: 5000			’alpha’: 0.1
			‘subsample’: 0.35

*Hyperparameters are optimal for lasso selected features, the feature set yielding highest average F1 macro score across the 4 models.

	Average F1 Macro Score Across the 4 models
sfs	57.59%
lasso	63.73%
rf	62.12%

Evaluation

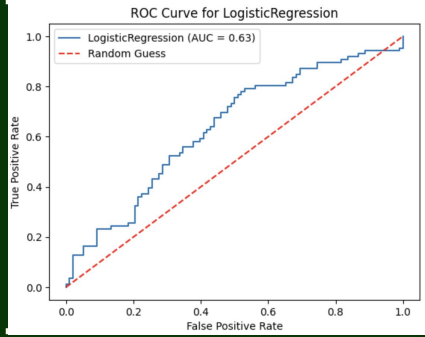
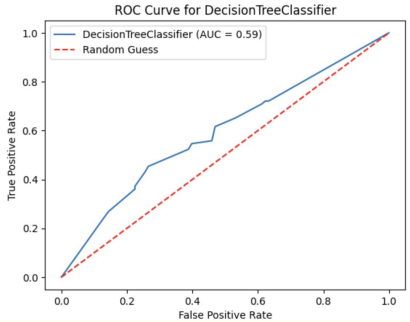
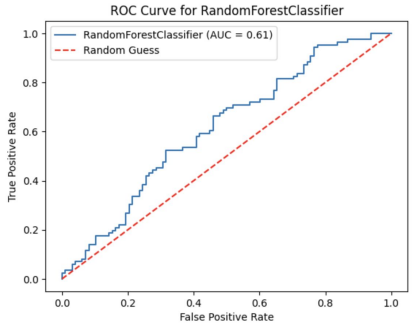
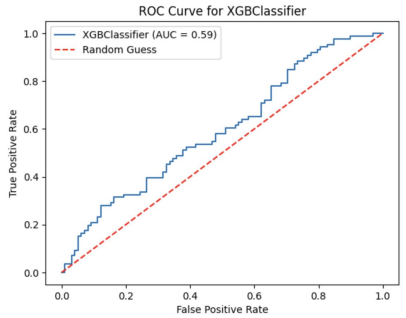
04

Test Performance Metrics

	Logistic Regression	Decision Tree	Random Forest	XGBoost
Accuracy	59.2%	**59.8%**	58.2%	57.1%
Precision	**59.0%**	58.8%	57.6%	54.7%
Recall	41.9%	46.5%	39.5%	**47.7%**
F1-score	49.0%	51.9%	46.9%	**50.9%**
ROC-AUC Score	**63.4%**	59.2%	61.2%	59.4%

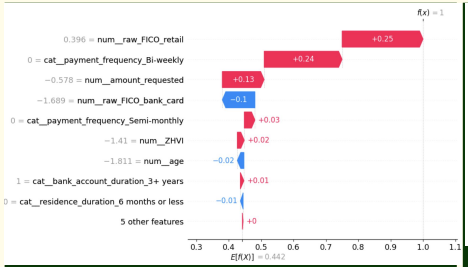
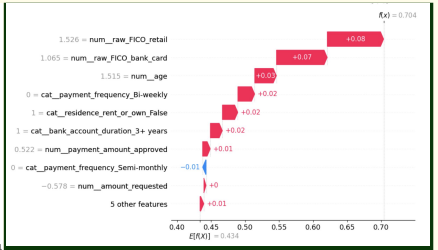
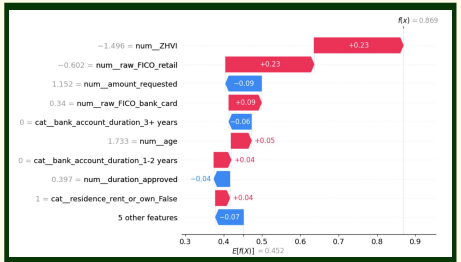
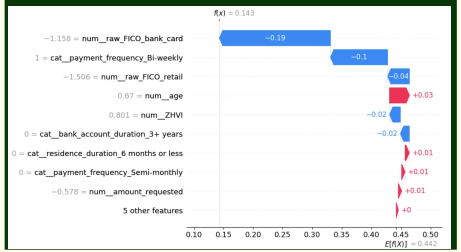
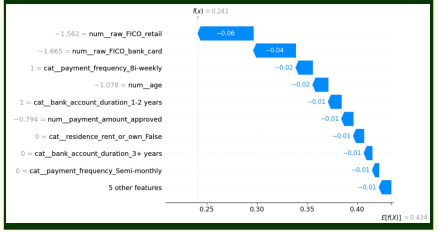
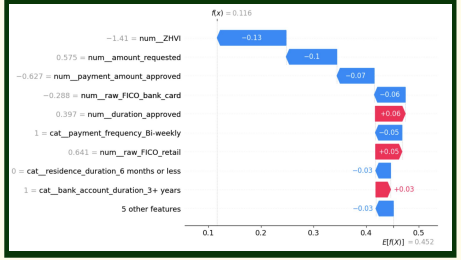
*Models fit to the 70% training set are evaluated on the 30% test set along the five metrics above. Models are trained on top features chosen from Lasso.

ROC Curves Evaluated on Test Set

Logistic Regression	Decision Tree	Random Forest	XGBoost
			

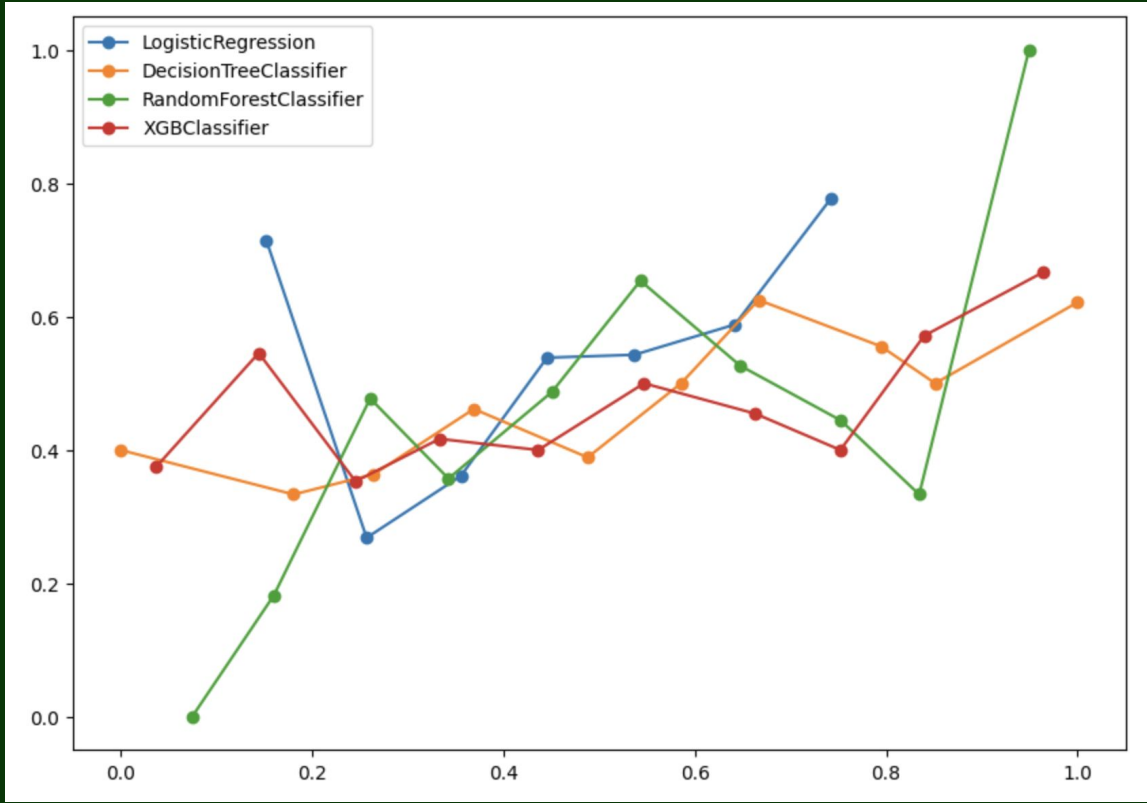
To promote economic equity through broader loan access, recall is the evaluation metric used for model comparison. Based on the classification performance metrics and ROC Curves, Logistic Regression is the top models in its ability to discriminate between “Good” and “Bad” loans. By recall, XGBoost is the best performing model at 47.7%, but overall is unable to reliably predict “Good” loan performances correctly. XGBoost may be underperforming because of the lack of well-dispersed loan data distributions; customers have features that cluster around similar values, such as loan duration or number of payments.

Shapley Value Plots (Customer)

Logistic Regression	Decision Tree	Random Forest	XGBoost
			
			

*Features are normalized, so numerical features have an average of 0.

Model Calibration



Discussion

05

Future Work

Data Integrity
- Validate customer location using bank routing numbers.

Feature Engineering
- Extract bank type (e.g., is the bank is a credit union) from bank routing number

Modeling Decisions
- Rerunning models with the worst or best performing loan in a 6 month period as the target variable can provide financial stress tests on worst-case or best-case scenarios of customer borrowing tendencies.
- Modeling was done on 612 customers, which is a fairly small sample. Also, the loan data is taken from a snapshot in 2010/2011, which is likely not representative of loan applicants present-day.

External Data Fusion
- Although other data sources were considered, such as natural gas usage from zip codes in 2010 ”Natural Gas Consumption by Zip Code”, zip codes did not give good coverage on the provided loan applicants.

Future Analyses
- For downstream adhoc analytics tasks, further visualizations can be made on discretized versions of variables such as Age into brackets to answer targeted business questions.
- Models with more parameters may be trained on gpu. For more exhaustive hyperparameter searches, a bayesian parameter tuning framework such as hyperopt may be leveraged on gpu clusters (e.g., Databricks Compute Cluster).