Stat 451 Final Project

Maya, Grace D, Meg, Grace S, Sowmya

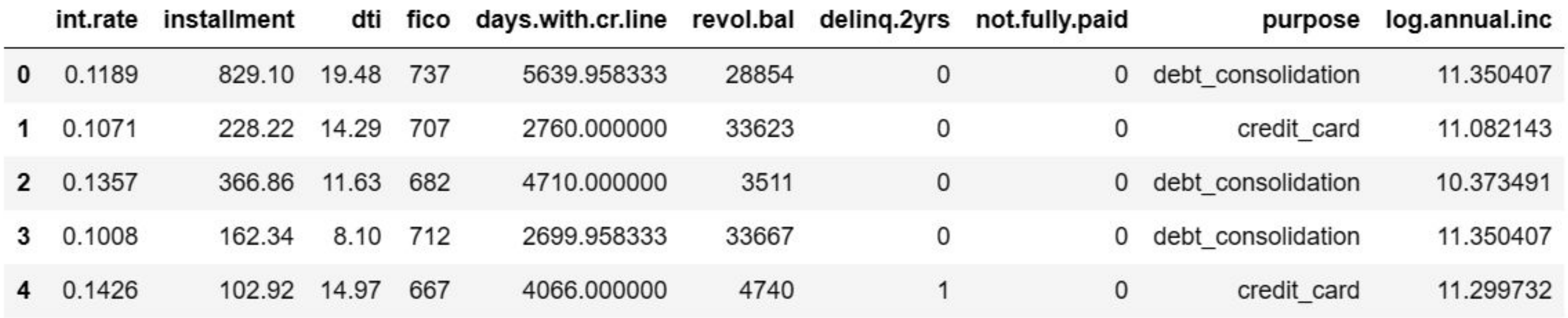
Loan Approval Project

Question(s):

- How can we classify whether a person will pay back their loan based on various features like FICO score, debt-to-income ratio, and interest rate?
- What features are important in determining whether a person will pay back their loan?

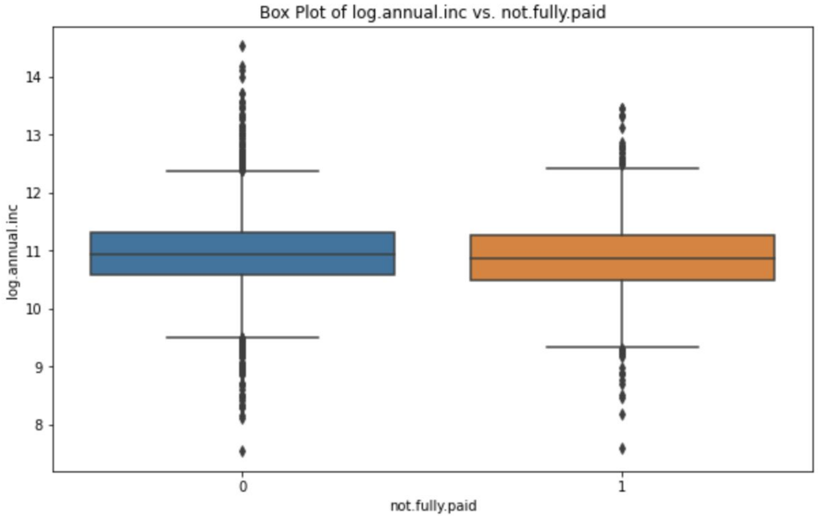
Kaggle lending dataset(2007-2010)

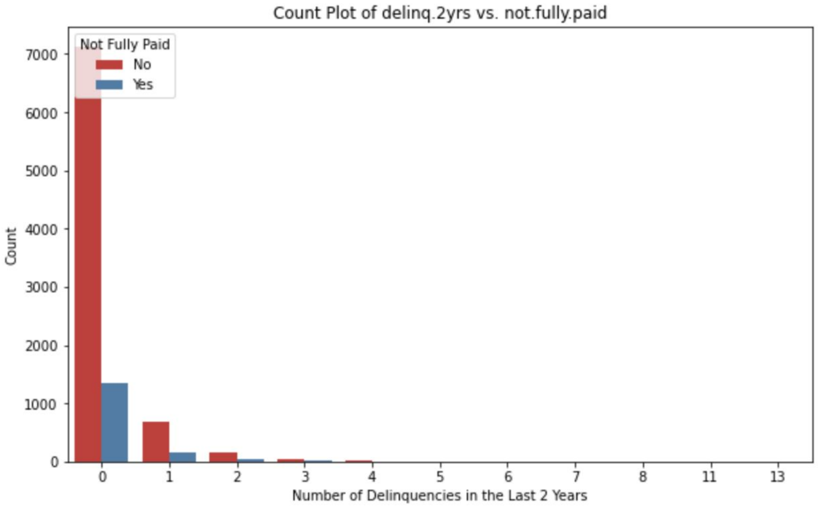
Variable we
are predicting



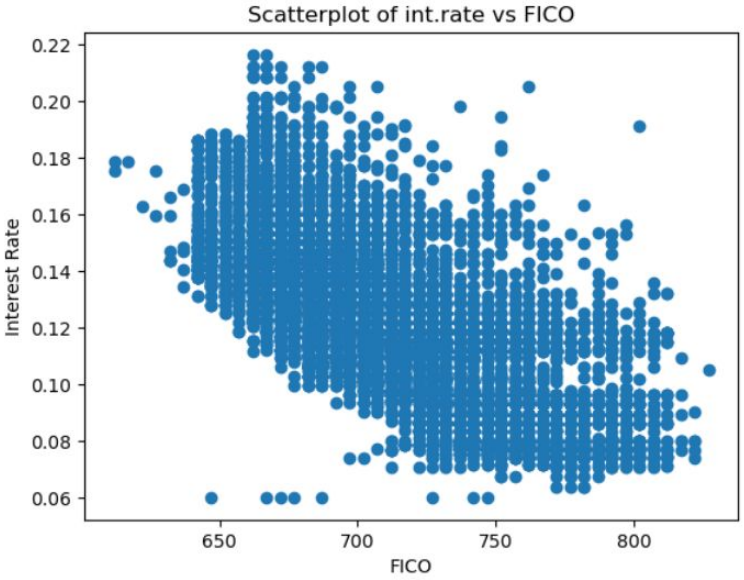
	int.rate	installment	dti	fico	days.with.cr.line	revol.bal	delinq.2yrs	not.fully.paid	purpose	log.annual.inc
0	0.1189	829.10	19.48	737	5639.958333	28854	0	0	debt_consolidation	11.350407
1	0.1071	228.22	14.29	707	2760.000000	33623	0	0	credit_card	11.082143
2	0.1357	366.86	11.63	682	4710.000000	3511	0	0	debt_consolidation	10.373491
3	0.1008	162.34	8.10	712	2699.958333	33667	0	0	debt_consolidation	11.350407
4	0.1426	102.92	14.97	667	4066.000000	4740	1	0	credit_card	11.299732

Data Exploration: Annual Income and Delinquency

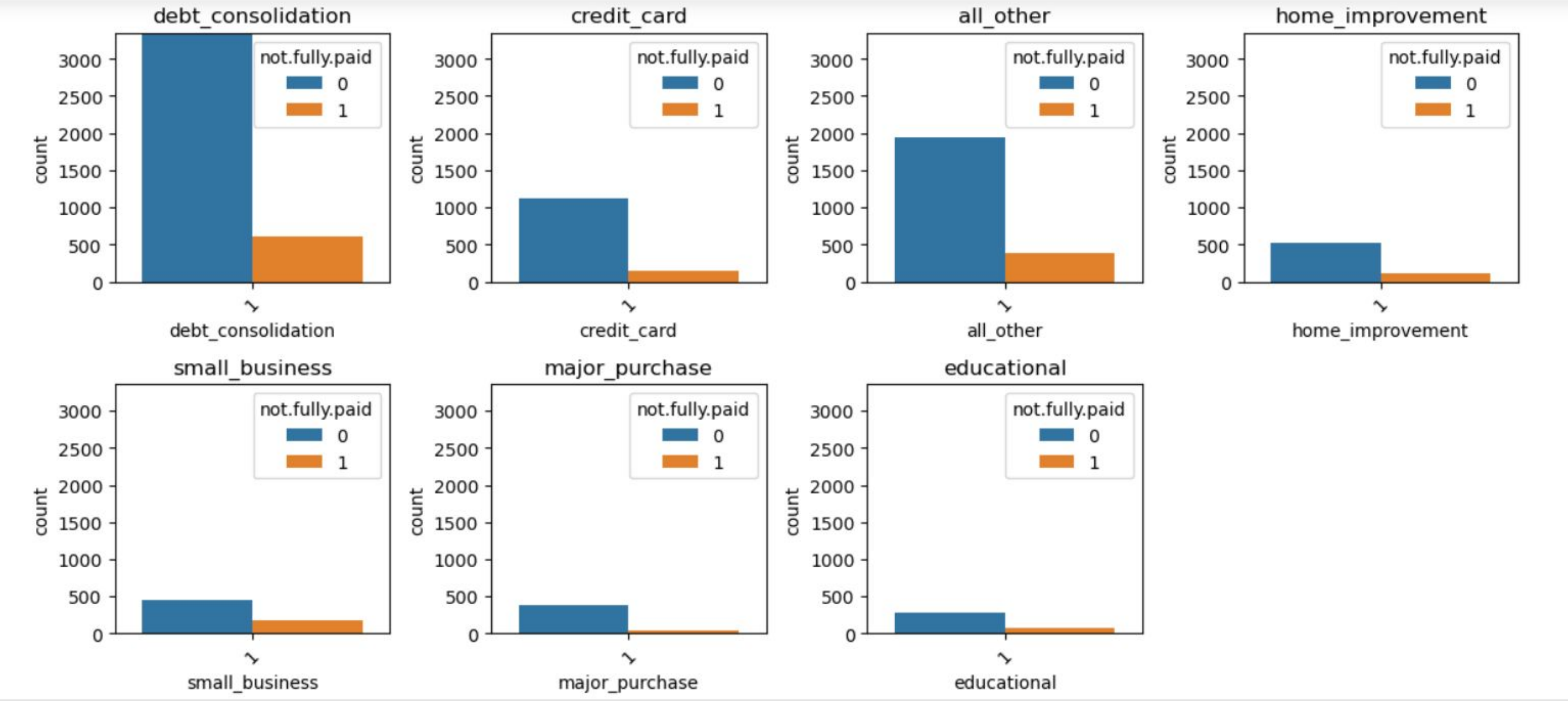




Data Exploration

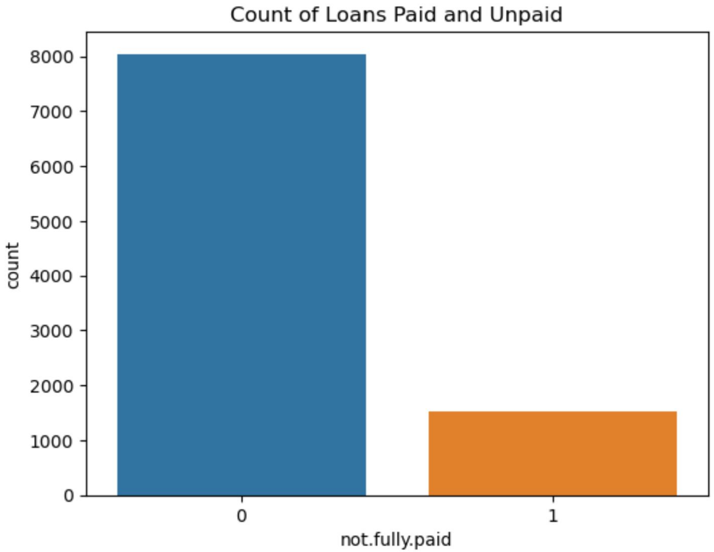


Loan Purposes



Paid vs. Unpaid Loans

- 83.99% of the loans from our data were fully paid back
- Guessing “0” every time yields an 83.99% accuracy rate



Feature Engineering

- Used One Hot Encoding to transform the purpose variable
- Used min-max normalization (for KNN classification)
- Tried to use feature selection with SelectKBest and f_classif to reduce overfitting(did not keep this as this came at the cost of a low accuracy)

Simple Model DecisionTreeClassifier()

Accuracy score of training data: 1

Accuracy score of validation data: 0.7537

Overfitting because training score > validation score

Methods

- Used hyperparameter tuning to find the best classifier/parameter pairing(s)
  - Added higher weights to the 1 class (did not use this as it made accuracy worse)
  - Varied log regression decision threshold(did not use this as it made accuracy worse)

Outcomes: Best classifiers and parameters

Logistic Regression

- parameters of C = .01, 10, 1000 all yielded the same accuracy

Decision Tree Classifier

- parameters of criterion entropy with max depth 1 and 2 yielded the same accuracy

Random Forest Classifier

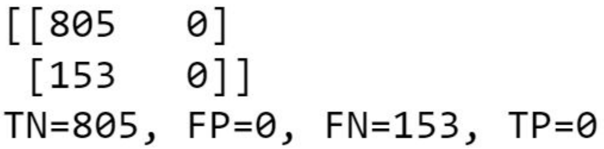
- parameter of max_depth = 11

The best accuracy score on validation data through hyperparameter tuning was 0.83925

Scores for Different Models on Test Data

Decision Tree

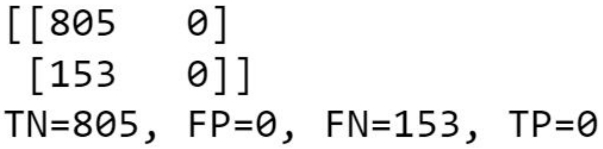
- accuracy = 0.8403



```
[[805   0]
 [153   0]]
TN=805, FP=0, FN=153, TP=0
```

Log Regression

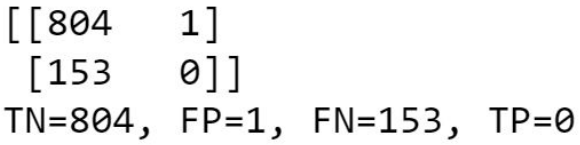
- accuracy = 0.8403



```
[[805   0]
 [153   0]]
TN=805, FP=0, FN=153, TP=0
```

Random Forest

- accuracy = 0.8392



```
[[804   1]
 [153   0]]
TN=804, FP=1, FN=153, TP=0
```

Assessing Model

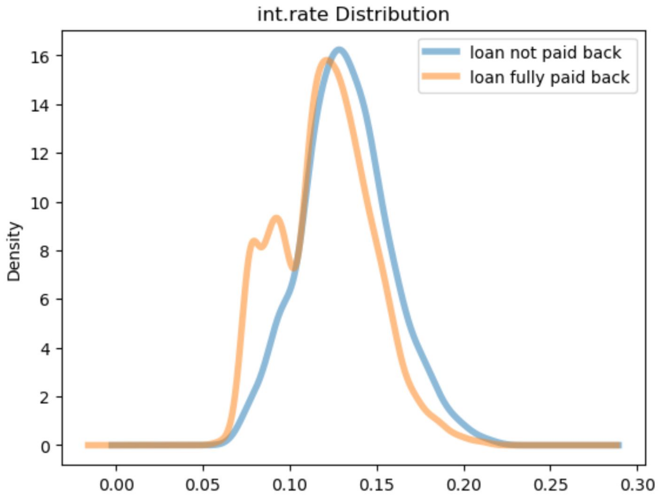
Used feature importance/analysis to find the most useful variables

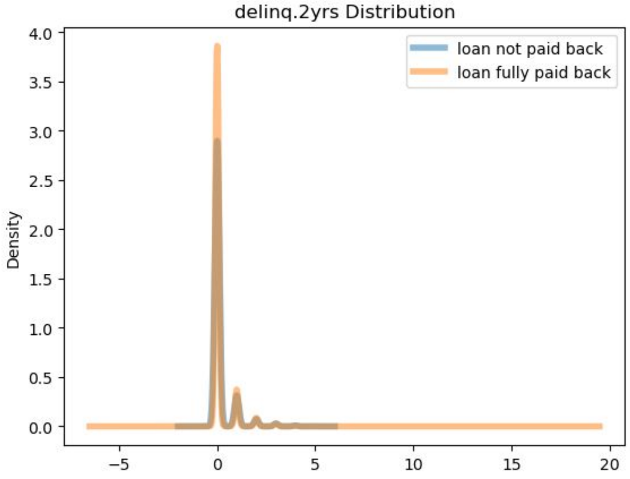
- Lasso
- Impurity based vs Permutation importance
    - Checked for collinearity

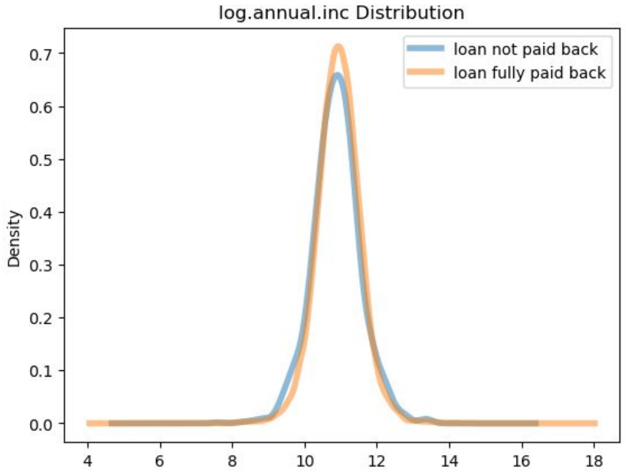
Outcomes: Lasso with Log Regression

Non zero variables (0.1 threshold):

- int.rate
- log.annual.inc
- delinq.2yrs
- And every purpose column

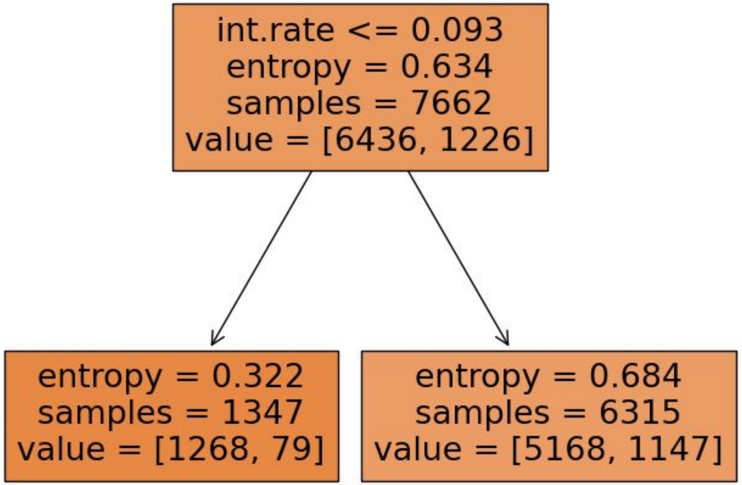






Decision Tree Plot

```
Accuracy on training data is clf.score(X, y)=0.8399457089162664.
```

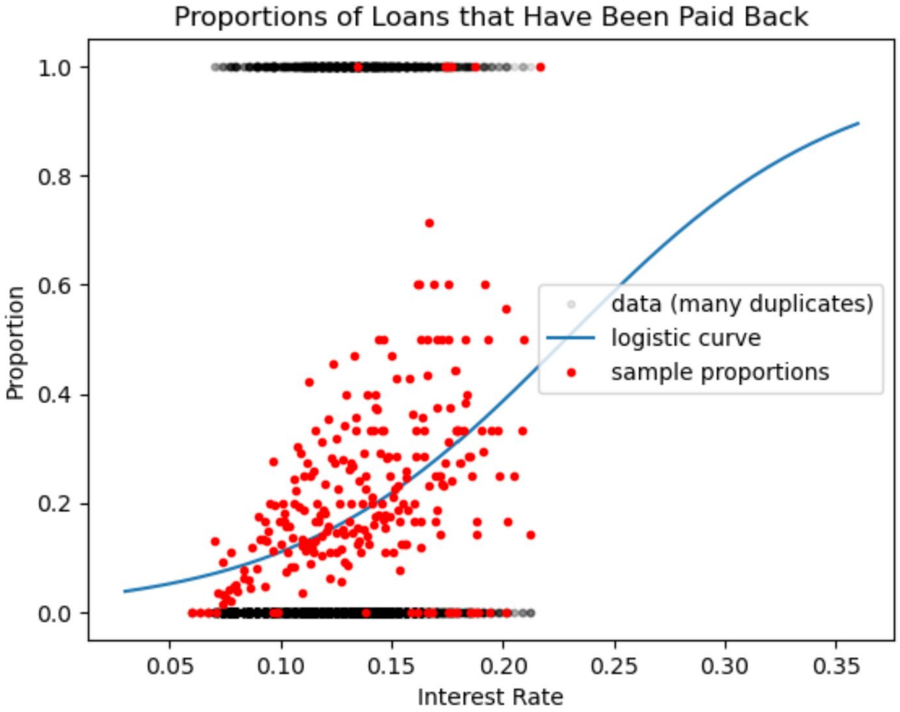


int.rate <= 0.093
entropy = 0.634
samples = 7662
value = [6436, 1226]

entropy = 0.322
samples = 1347
value = [1268, 79]

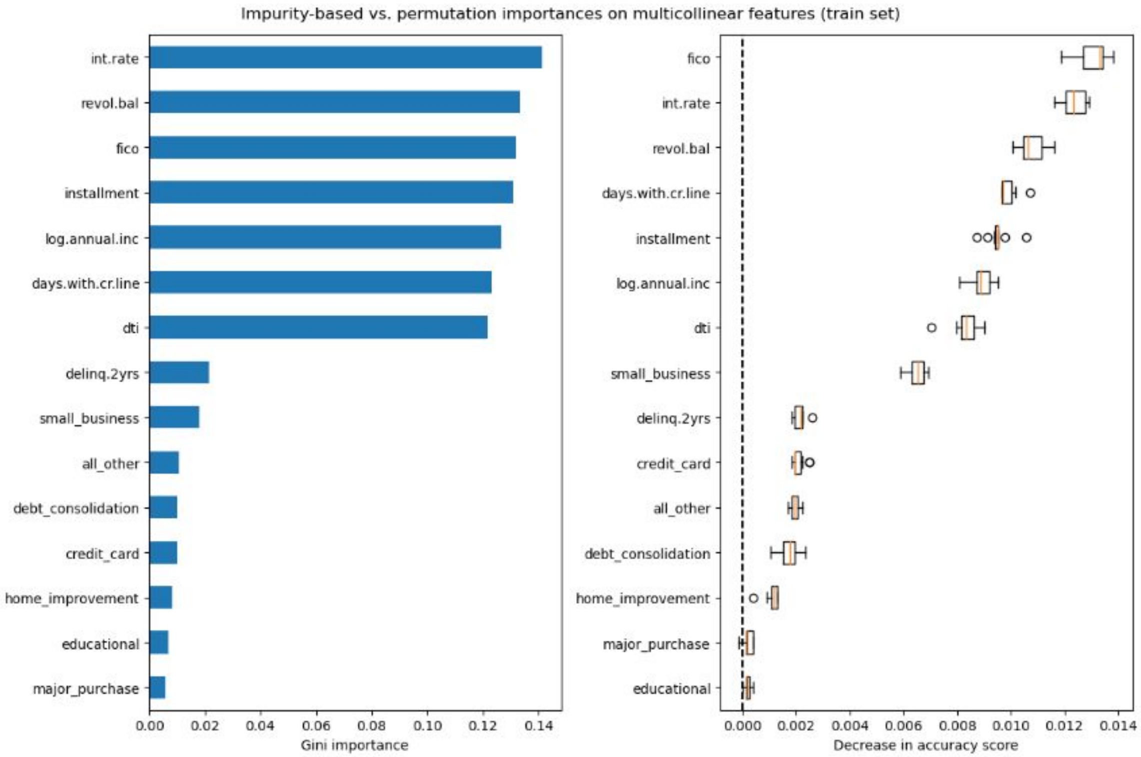
entropy = 0.684
samples = 6315
value = [5168, 1147]

Logistic Regression Graph



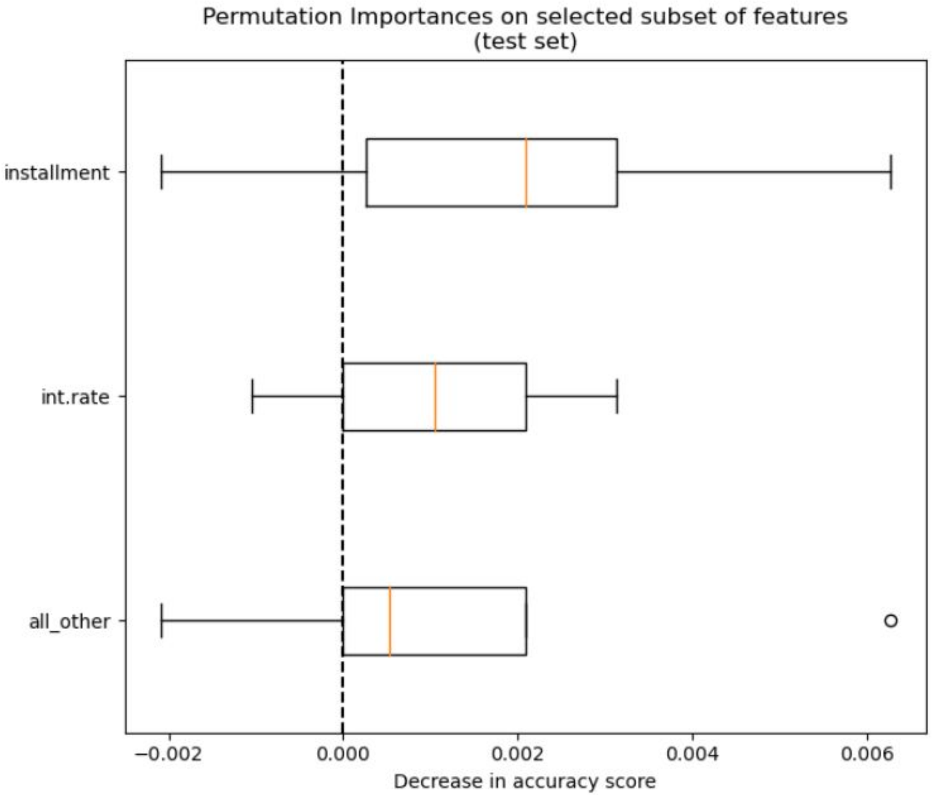
Random Forest Classifier Feature Importance

https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html#sphx-glr-auto-examples-inspection-plot-permutation-importance-multicollinear-py



Random Forest Classifier Feature Importance

https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html#sphx-glr-auto-examples-inspection-plot-permutation-importance-multicollinear-py



Conclusion

- Maybe predicting 0 is best until we find a variable that is more informative to correctly predict not paying a loan.
- How could f1 score(combination of recall and precision) help evaluate the best model instead of accuracy?