AUGUST 2024

LOAN APPROVAL PREDICTION MODEL

A SUPERVISED MACHINE LEARNING CLASSIFICATION PROJECT

Prepared by: Me!
DTSA 5509 Final Project

As part of the MS-DS program
University of Colorado at Boulder

01

ABOUT THIS PROJECT

Create a machine learning model which can predict the approval status of a loan.

WHY?

Buy a house

Buy a car

Start a business

Scalable solution to save time and money!

02

THE DATA

Sharma, A. (2021). *Loan approval prediction dataset*. Kaggle.

https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset. License: MIT

License. https://www.mit.edu/~amini/LICENSE.md

03

THE DATA

- **Employment status**
- **Number of dependents**
- **Education**
- **Income**
- **Loan amount**
- **Credit score**
- **Assets**
- **Approval status - the target**

04

- **SimpleImputer**
- **ColumnTransformer**
- **StandardScaler and OneHotEncoder**
- **Pipeline**
- **SVC**
- **kNN**
- **RandomForest**

SCIKIT LEARN

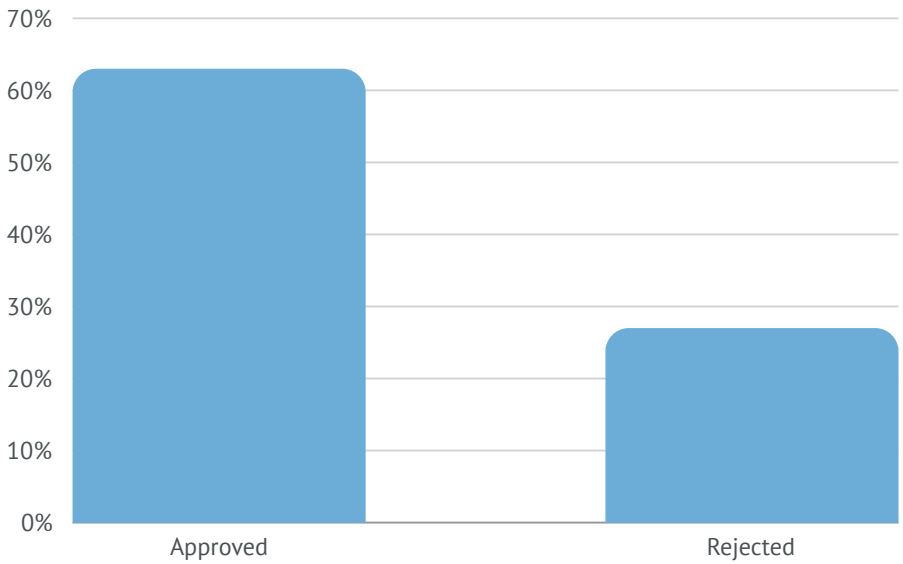
05

EDA: DISTRIBUTIONS

Target Data

- Imbalanced target data
- Source of bias

-> class weights
-> resampling, ensembling
-> evaluation metrics

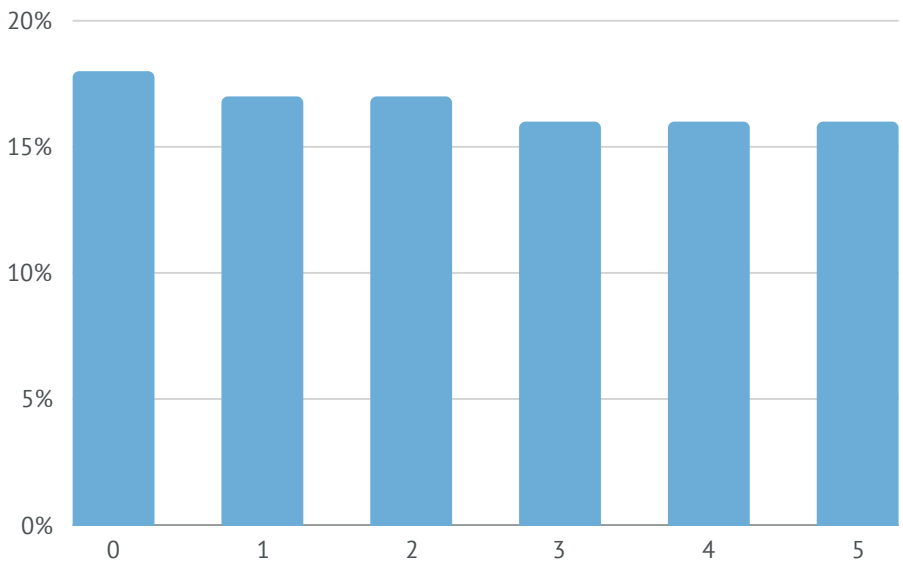


06

EDA: DISTRIBUTIONS

Categorical Features

- Suspected they weren't very important
- Eg: dependents

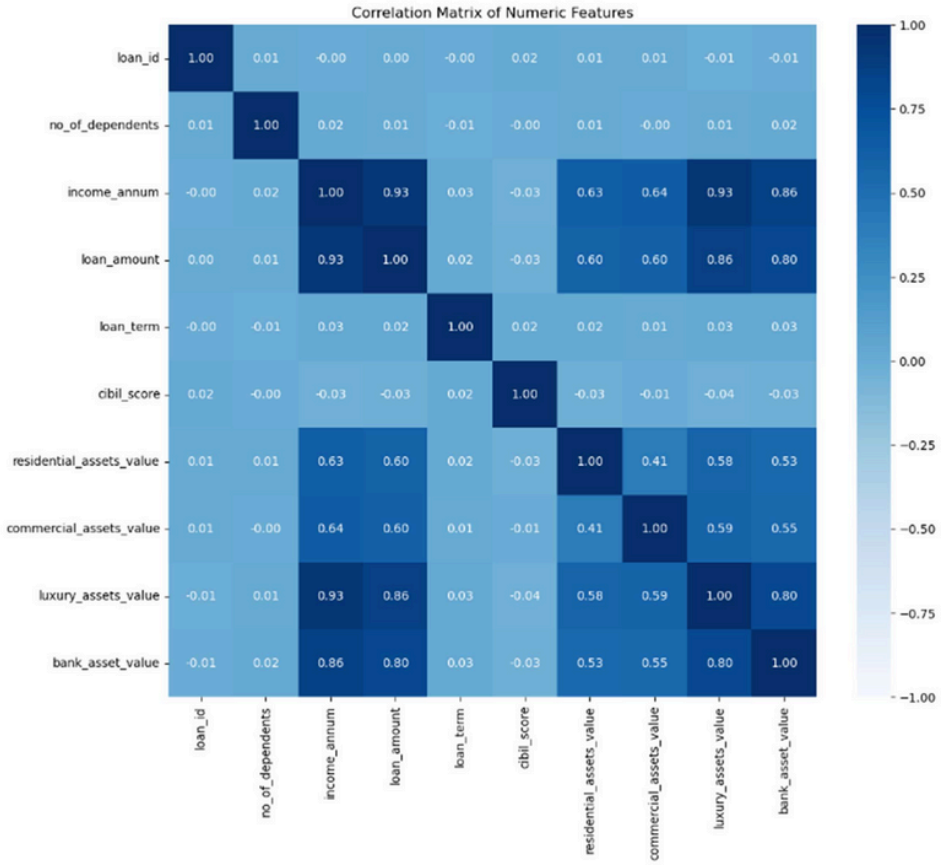


07

EDA: DISTRIBUTIONS

Numerical Features

- Strong correlations
- Need to assess multicollinearity



08

FEATURE SELECTION

Categorical

- **Chi-squared < 1**
- **p-value > alpha = 0.05**
- **As suspected, not significant**

Feature	Chi-Sqr Score	p-value
education	0.209	0.647
self_employed	0.019	0.889
no_of_dependents	0.386	0.534

09

FEATURE SELECTION

Numerical

- **Variance Inflation Factor (VIF)**
- **High VIF: 5 to 10**
- **Income VIF = 74!**
- **Removed features until...**

Feature	VIF
loan_amount	7.24
credit_score	4.93
loan_term	3.90
residential_asset	3.60
commercial_asset	3.60

10

MODELING: ITERATIVE PROCESS

1. type_of_feat = [ ]
2. specify transformers
3. preprocessor to put features through transformers
4. list of models
    a. pipeline
    b. cross-validate
    c. results
5. compare results

MODELING: ROUND 1, ALL FEATURES

Model	Mean Train Accuracy	Mean Validation Accuracy
Dummy	0.6310	0.6310
SVM	0.950	0.935
kNN	0.933	0.890
Random Forest	1.000	0.975

MODELING: ROUND 3, REDUCED FEATURES

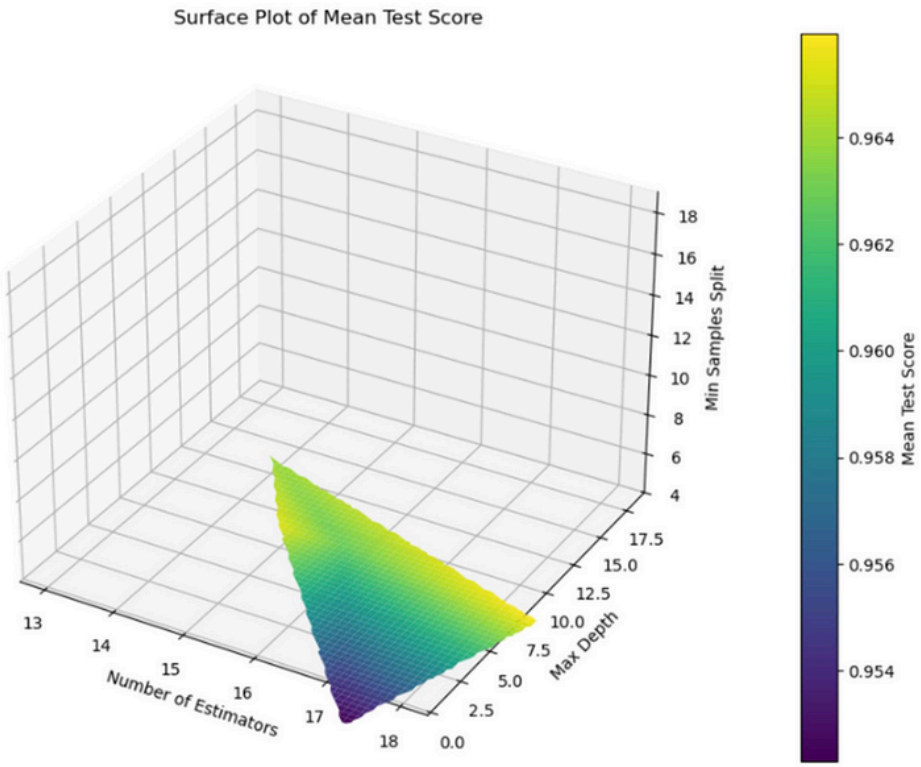
Model	Mean Train Accuracy	Mean Validation Accuracy
Dummy	0.6310	0.6310
SVM	0.943	0.935
kNN	0.957	0.925
Random Forest	**1.000**	**0.9672**

13

HYPERPARAMETER TUNING

Parameter grid search

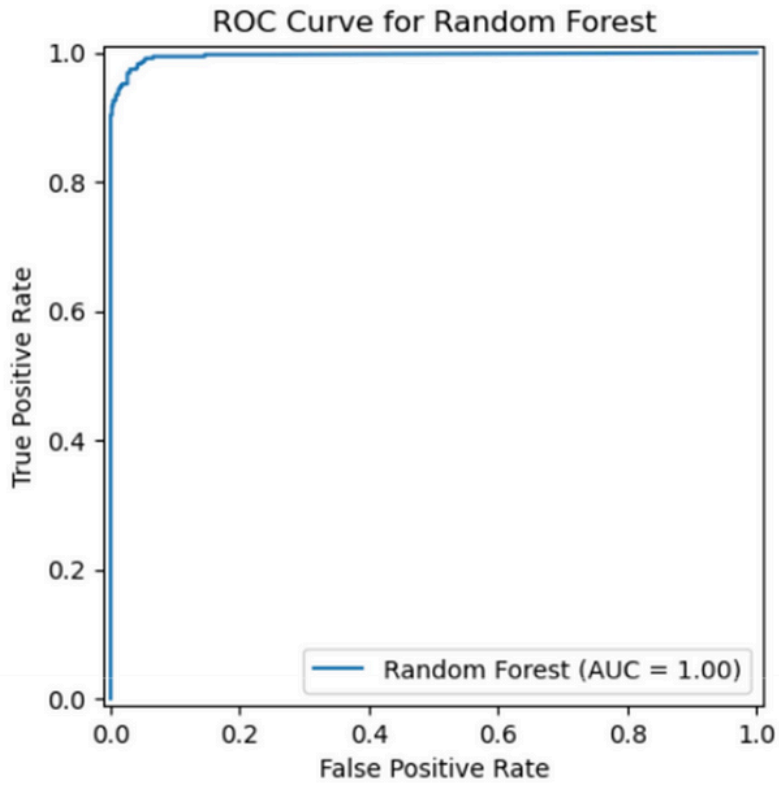
- **number of estimators = 18**
- **max depth = 12**
- **min samples split = 19**



14

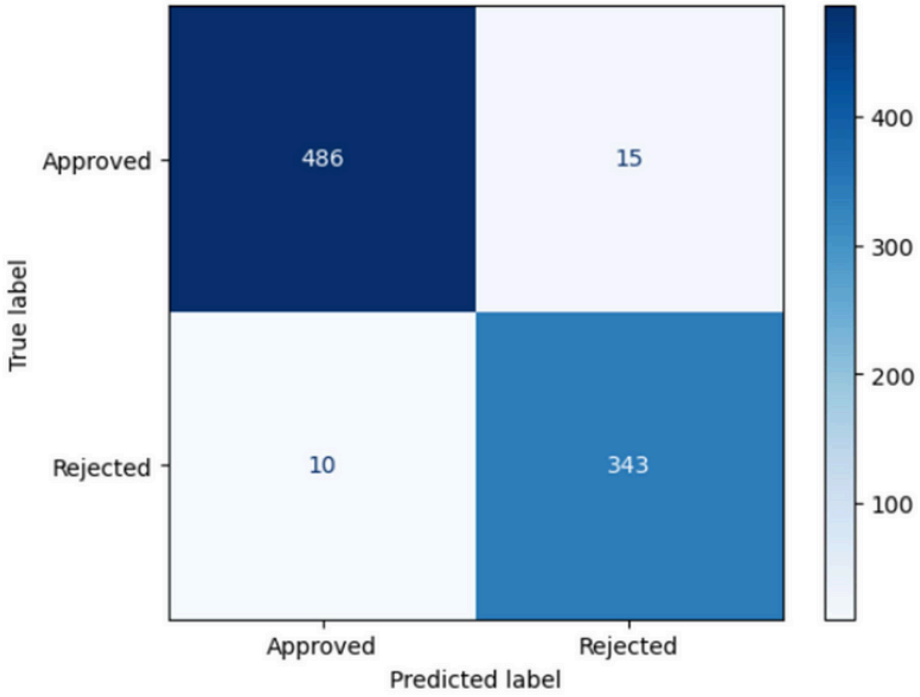
MODEL EVALUATION

rf_test_score = rf_search.score(X_test, y_test)
rf_test_score = 0.9707
AUC = 0.9959



AUC: 0.9959

MODEL EVALUATION



	Precision	Recall	F1 Score
Approved	0.98	0.97	0.97
Rejected	0.96	0.97	0.96

16

DISCUSSION

> *"The dataset and model are quite simple, and used for demonstration purposes only. This model should not be used to approve actual loans"*
>
> -ME

Further improvements:

- addressing class imbalance before the pipeline could have helped to improve the kNN model.
- statistical testing using paired t-tests on changes in the mean training and validation accuracies - significance?
- Ethical considerations - people are highly diverse! Minorities and marginalized persons may not be represented in the data.

17

CONCLUSION

> *"It's pretty good"*
>
> -ME