Loan Approval Analysis

Introduction

The primary goal of my project was to develop a predictive model which would increase the efficiency and accuracy of the process of obtaining approval for a home loan. Being able to quickly identify and approve customer loans based upon their application data has wide reaching impacts for both lenders and borrowers.

While the approval process is complex, and many factors – prior credit history, income, and loan amount for example – are taken into consideration before an individual receives disbursement of funds, there is inherent risk that a lender may approve a loan for someone who may ultimately cost this lender money. In fact, according to the Consumer Financial Protection Bureau (2023), the percentage of mortgages in the US that were between 30-89 days delinquent from 2008-2022 ranged between 4 to 1 percent.

For consumers, delays in loan approval can cause considerable disappointment in a highly competitive housing market. Comparing June 2022 to June 2023, there was a year over year decrease of 29 % of new homes listed (809,234 vs. 576,897). For the total number of homes for sale for the same months, reports indicate a 17% decrease in available inventory (1,113,438 vs. 921,004). To compound matters, the median number of days on the real estate market for this June was 29 days before an offer was accepted (Redfin, 2023). What this can mean is that if there is anything that could be causing a delay with the loan approval process, a consumer may find themselves looking for a home for quite some time.

We hope that this modeling will offer financial institutions a new approach for establishing loan eligibility, allowing them to provide better and faster services to more potential customers, creating value for both lenders and borrowers alike.

Methods

Many datasets were reviewed during the early stages of the project, with particular emphasis on finding datasets of sufficient length and enough features to be able to draw meaningful conclusions. Ultimately, we landed on a dataset obtained via Kaggle which was comprised of 614 rows and 13 columns (Kumar, 2020) due to its high quality and reasonable number of features. These features consisted of expected information from loan applications including but not limited to gender, income, marital status, loan amount, and education. During exploratory data analysis, we identified both missing data and outliers (values more than 3 standard deviations from the mean for our numerical features) and removed these

as part of our data cleansing steps. Oversampling was employed to address data imbalance in the target variable. Dummy variables were generated and then the best features were selected to reduce the dimensionality from 465 to just 30 features. An 80/20 split was employed to create training and test datasets for modeling and both sets were normalized with RobustScaler.

As the project at its core is a binary classification problem – *is this loan approved or not* – we discussed suitable candidates which would work well for this purpose, making the decision to compare logistic regression and random forest models as shown below:

```
#Random Forest
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(random_state=42)
rf.fit(x_train_scaled, y_train)
```

```
#Logistic Regression
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import *

lr = LogisticRegression(max_iter=1000, solver='liblinear')
lr.fit(x_train_scaled, y_train)
```

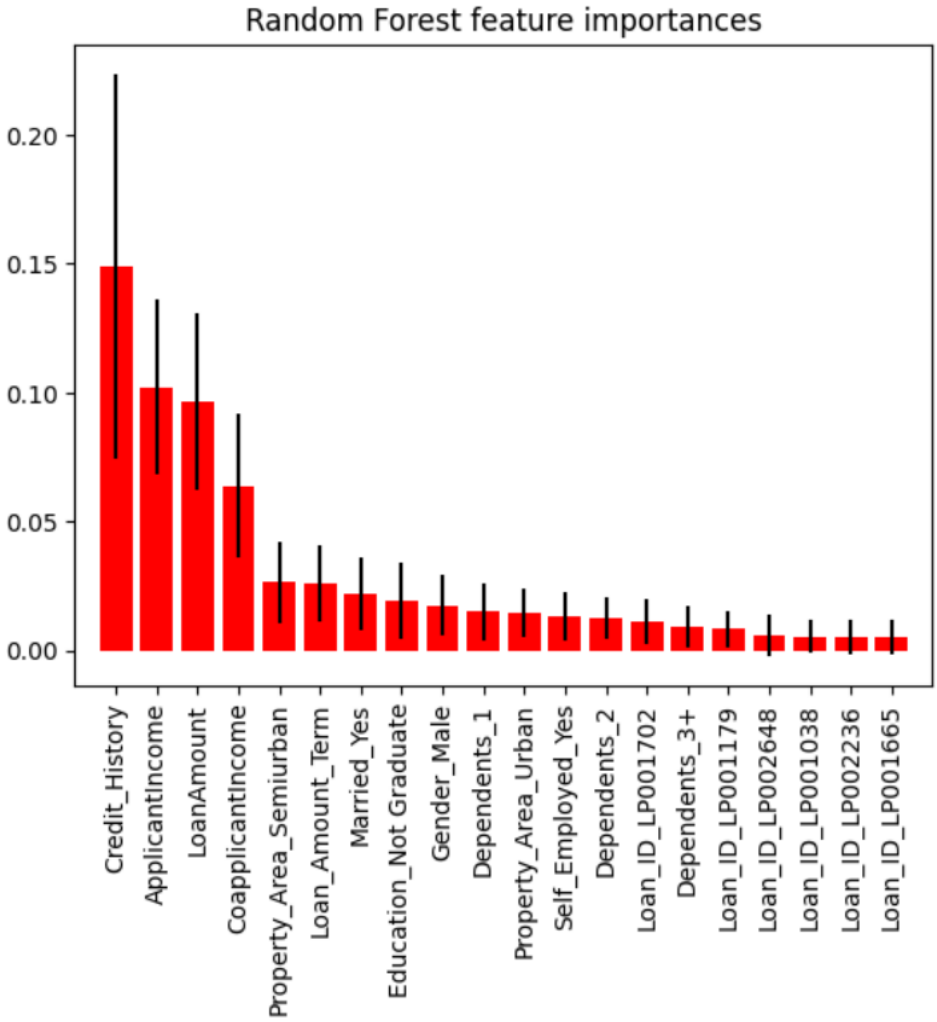
These models were evaluated using accuracy, precision, recall and F1-score. While the initial results were likely more than suitable, hyperparameter tuning with grid search and cross validation was nonetheless preformed to enhance the modeling. The predictability of these subsequent models was reevaluated using the same measures as prior untuned versions.

Results

Initial results on the untuned models suggested that the Random Forest Classifier could likely be a better model for the purposes of predicting loan approvals, with accuracy, recall, precision, and F1-scores all outperforming those of the logistic regression.

	Model	Accuracy	Precision	Recall	F1 score
1	Random Forest	94.0	100.0	87.0	93.0
0	Logistic Regression	80.0	78.0	84.0	81.0

An analysis of the feature importance in the random forest model showed that 4 features were considerably important in predicting loan approval – credit history, applicant income, loan amount, and coapplicant income.



However, after hyperparameter tuning was completed, we found that the logistic regression proved to be a better model for predicting loan analysis, with the best logistic regression model possessing accuracy, precision, recall, and F1- scores at 98% while the Random Forest Classifier's accuracy and F1-scores dropped to 91% each. Below are the hyperparameters and the accuracy measures for our respective tuned models.

Logistic Regression - After Parameter Tuning

```
#Logistics regression parameter tuning

from sklearn.model_selection import GridSearchCV

params = {

    'C':[1,5, 10, 20],
    'penalty':['l1', 'l2']
```

```
}

lr_grid = GridSearchCV(lr, params, n_jobs=-1)
lr_grid.fit(x_train_scaled, y_train)

print(lr_grid.best_params_)
print(lr_grid.best_score_)
```

```
{'C': 5, 'penalty': 'l1'}
0.9069306930693071
```

Random Forest Model - After Parameter Tuning

```
params = {
    'max_depth':[2,3,4,10,20],
    'n_estimators':[50, 100, 120]
}


rf_grid = GridSearchCV(rf, params, cv=2, n_jobs=-1)
rf_grid.fit(x_train_scaled, y_train)

print(rf_grid.best_params_)
print(rf_grid.best_score_)
```

```
{'max_depth': 20, 'n_estimators': 100}
0.7921607378129117
```

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	98.0	98.0	98.0	98.0
1	Random Forest	91.0	88.0	94.0	91.1

Conclusion

The high accuracy and F1 values both before hyperparameter tuning suggest that there may be some degree of overfitting which has occurred with our modelling. We, however, believe the L1 and C=5 hyperparameters of the tuned logistic regression model mean that we have adjusted for these issues in a way that successfully addresses these issues of fit and feel confident in suggesting that logistic regression is the most reasonable predictive model for loan approval.

Given time constraints and limitations imposed by the size of the dataset, future research opportunities should be focused on expanding both the number of unique loan applications upon which the model is trained and adding additional features which would normally be derivable from application data, taking

great care to not create a system which either through carelessness or malicious intent denies loans along the lines of a federally protected class.

References

Downloadable Housing Market Data. (2023). [Dataset]. Redfin. https://www.redfin.com/news/data-center/

Kumar, G. R. (2020). *Loan Approval Data Set* (Version 2) [Dataset]. https://www.kaggle.com/datasets/granjithkumar/loan-approval-data-set

Mortgages 30-89 days delinquent. Consumer Financial Protection Bureau. (2023, June). https://www.consumerfinance.gov/data-research/mortgage-performance-trends/mortgages-30-89-days-delinquent/#mp-line-chart-container