Parking Tickets in Chicago

Predicting Payment

Agenda

Overview - Data - Baseline - Final Models - Next Steps

Overview

- Data from [ProPublica]
- City of Chicago Parking Tickets
- Passenger Vehicles only
- Multiple decades and has over 50 million observations
- Only analyzing first million

- Potential to help allocate city resources/increase revenue

Data

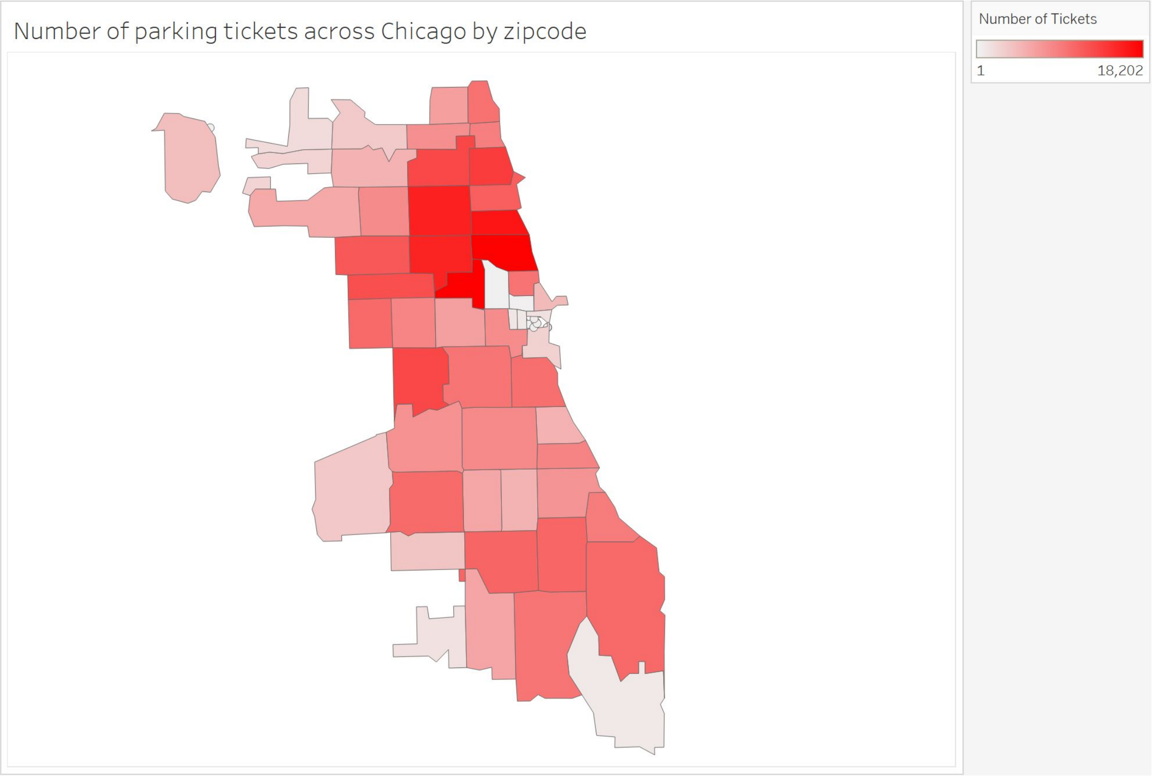
Target: Payment Status

- Paid if paid
- Not Paid if: Dismissed, Unpaid, Hearing Required, Notice Sent, or Bankruptcy

Features for Focus

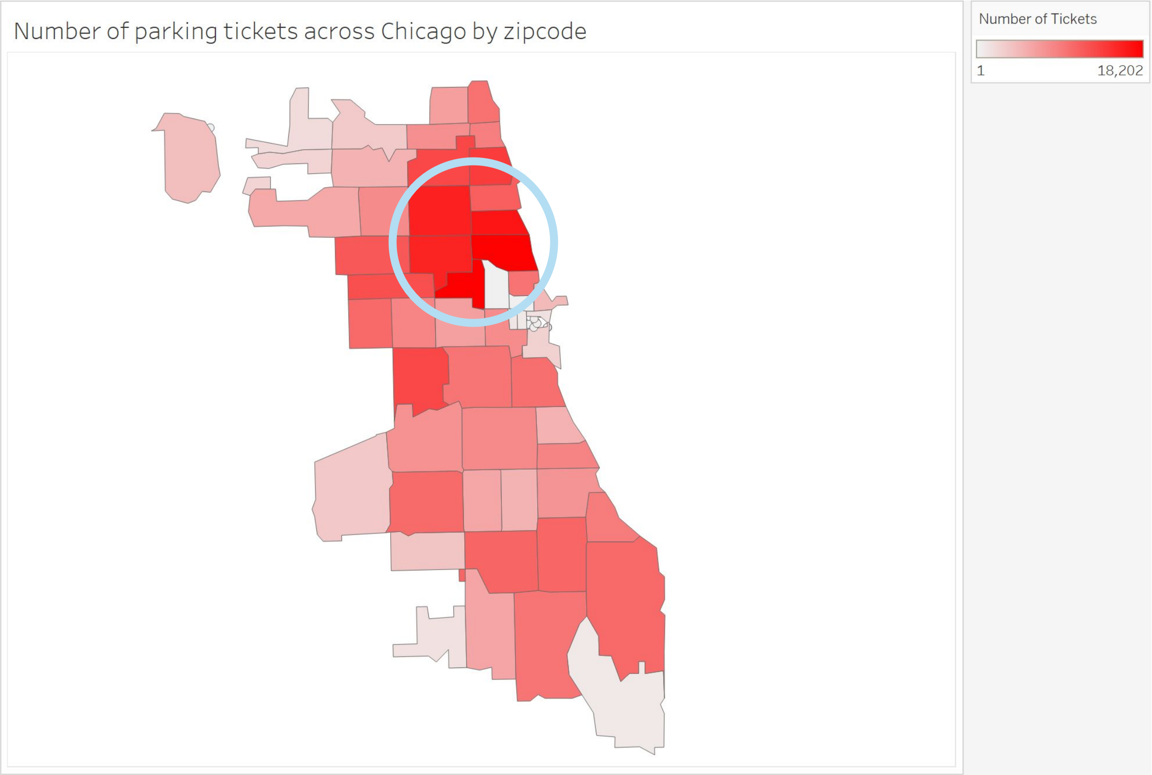
- License Plate State
- Geolocation (Latitude/Longitude)
- Fine amount ($$$)
- Violation code (what is the ticket for?)
- Count of license plate appearing in data

EDA

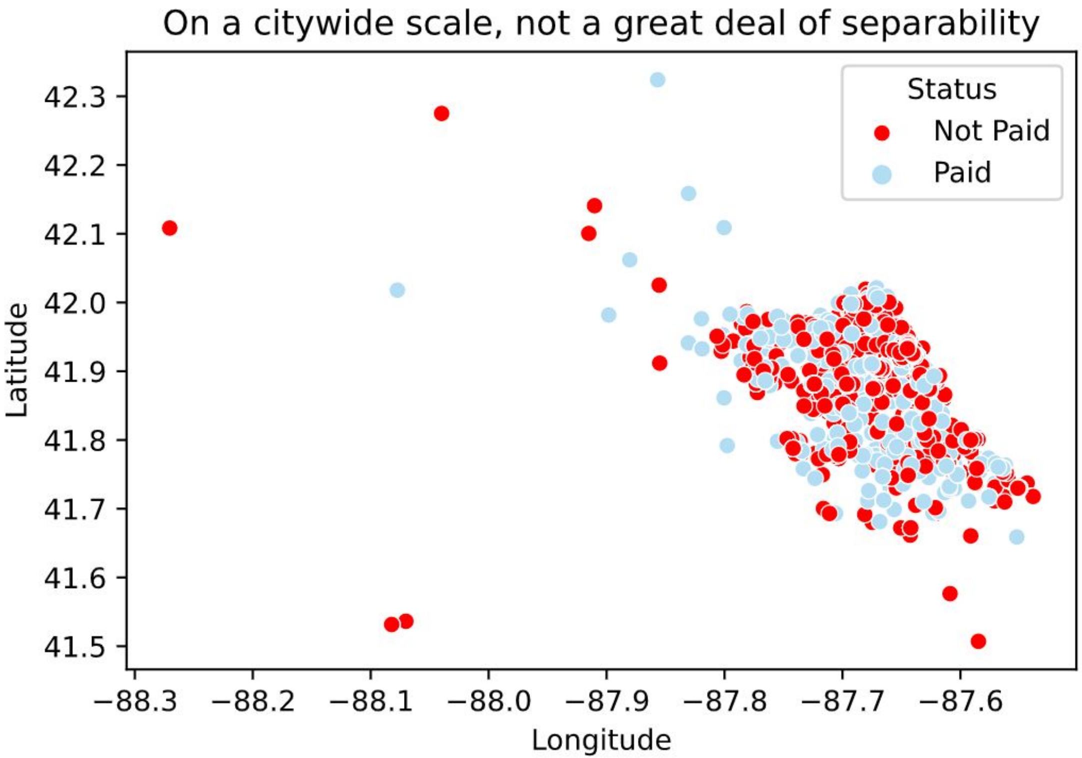


Number of parking tickets across Chicago by zipcode
Number of Tickets
1
18,202

EDA


Number of parking tickets across Chicago by zipcode
Number of Tickets
1
18,202

EDA


On a citywide scale, not a great deal of separability
Status
Not Paid
Paid
Latitude
42.3
42.2
42.1
42.0
41.9
41.8
41.7
41.6
41.5
Longitude
−88.3
−88.2
−88.1
−88.0
−87.9
−87.8
−87.7
−87.6

Thoughts on classifying and potential errors

Imbalance of paid tickets to unpaid (two to one): Random Oversample Unpaid

The models in this project classify unpaid tickets as “positive”

Thoughts on classifying and potential errors

Imbalance of paid tickets to unpaid (two to one): Random Oversample Unpaid

The models in this project classify unpaid tickets as “positive”

False Positives: Tickets the model expects to be unpaid, but actually are paid

False Negatives: Tickets the model expects to be paid, but actually are unpaid

Thoughts on classifying and potential errors

Imbalance of paid tickets to unpaid (two to one): Random Oversample Unpaid

The models in this project classify unpaid tickets as “positive”

False Positives: Tickets the model expects to be unpaid, but actually are paid

False Negatives: Tickets the model expects to be paid, but actually are unpaid

Goal: To separate paid and unpaid tickets as cleanly as possible

Metric: Use AUC score and confirm with confusion matrix

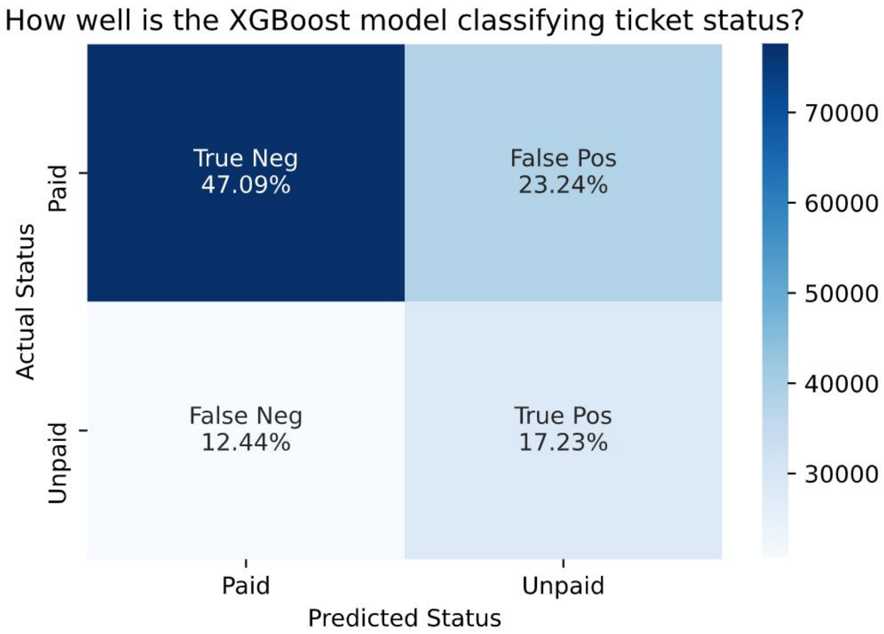
Baselining

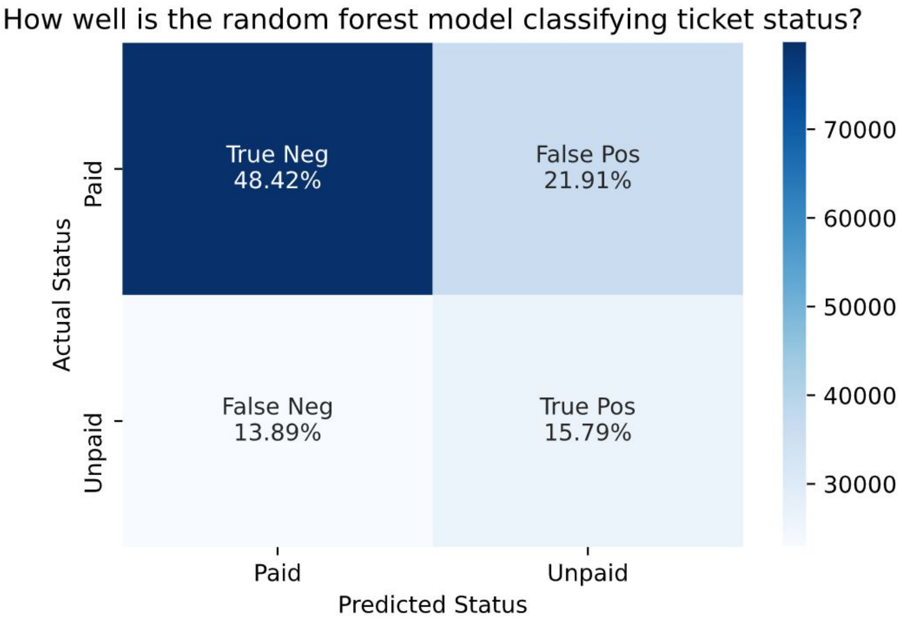
- Simple Logistic models based on each feature for smaller data sample
    - AUC scores near 50%, not much better than coin flip
- Simple kNN models similar to above
    - Performance (speed) is terrible on larger data sets, not worth waiting
- Use more features in training logistic, Random Forest, and XGBoost

The top performing models were

1) XGBoost
2) Random Forest

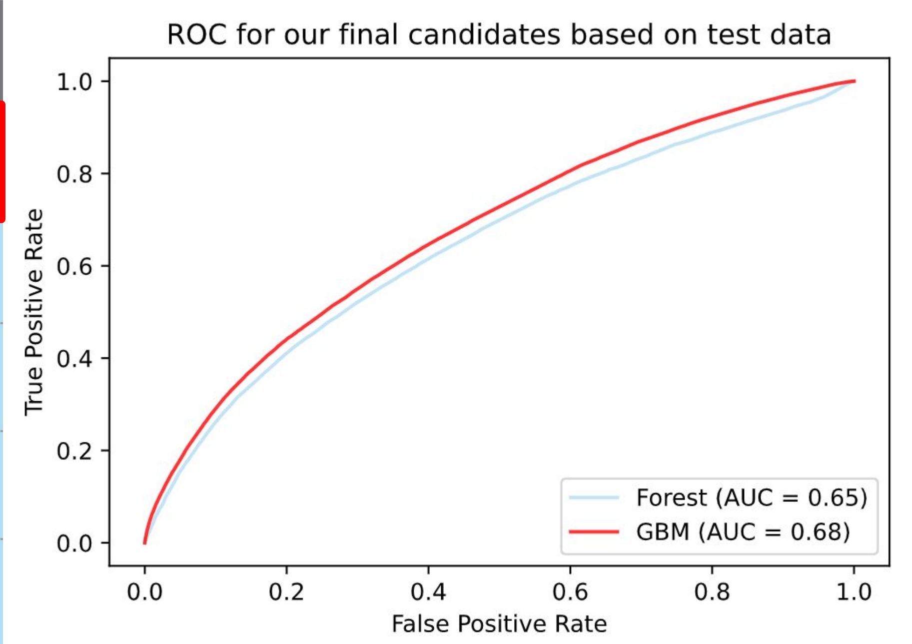
XGBoost and Random Forest similar on unseen data





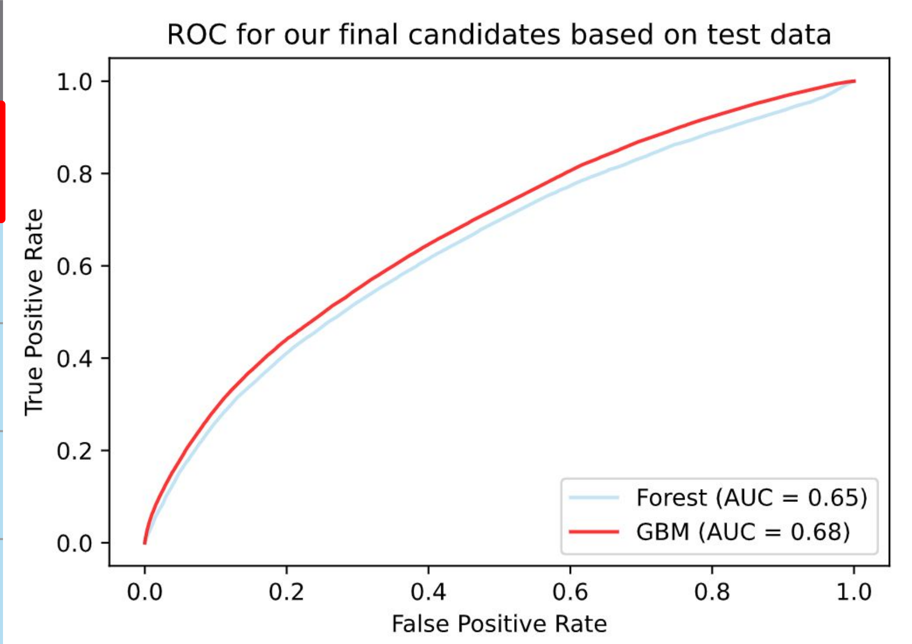
XGBoost generally outperforms Random Forest

Score	Random Forest	XGBoost
AUC	.65	.68



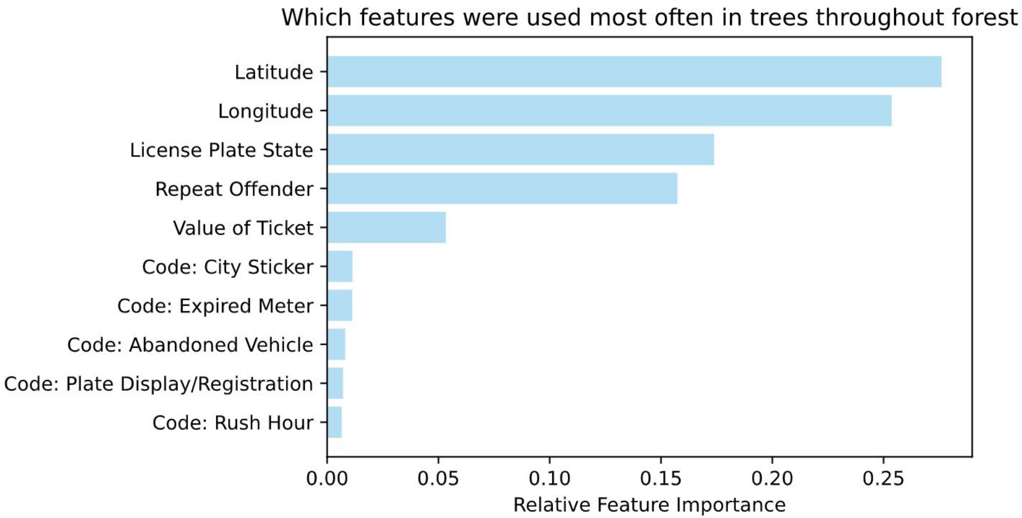
XGBoost generally outperforms Random Forest

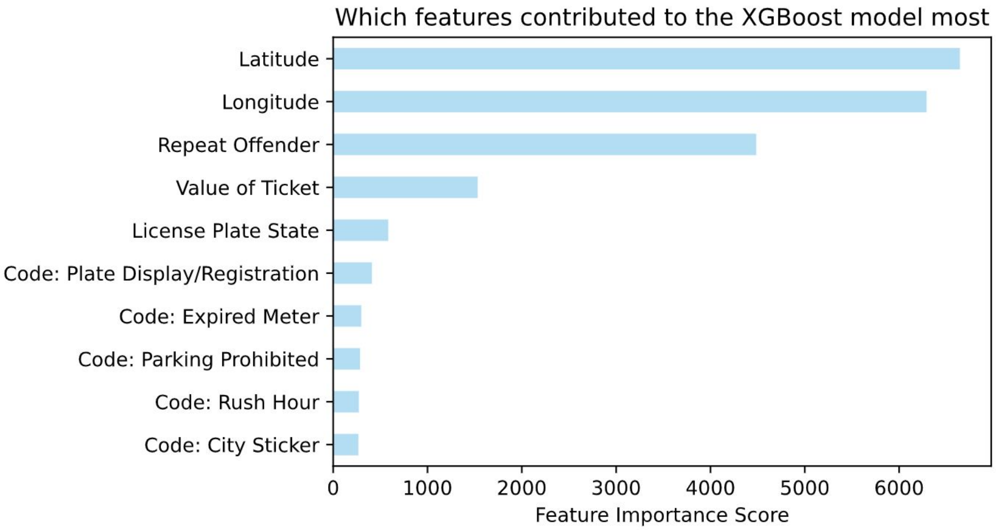
Score	Random Forest	XGBoost
AUC	.65	.68
Accuracy	.64	.64
Recall	.53	.58
Precision	.42	.43
F1	.47	.49



Feature Importance

- General agreement across models
- Ticket location used most
- Many violations, importance spread disaggregated





Potential Applications

Predict whether a given parking ticket will be paid or not paid

A) Since tickets are only written if infractions are found, the city must decide how to allocate employees. To generate more revenue, identify areas and tickets more likely to yield payment.

B) Since tickets are intended to be consequences, the city should be looking to monitor its citizens fairly, so no change should occur to ticket writing. However, it would be useful to know if the city could flag a ticket as being more likely to be delinquent and need following up.

Future Work

Use cloud computing to handle full dataset

Future Work

Use cloud computing to handle full dataset

Continue tuning hyperparameters to improve the model performance

Future Work

Use cloud computing to handle full dataset

Continue tuning hyperparameters to improve the model performance

Rather than paid or not paid, dig deeper:

> **Given these features, predicting if ticket goes to court**
> **Given ticket goes to court court cases, predicting judgements**
> **Given these features, predicting car seizure**

Future Work

Use cloud computing to handle full dataset

Continue tuning hyperparameters to improve the model performance

Rather than paid or not paid, predict:

Given these features, predicting if ticket goes to court
Given ticket goes to court court cases, predicting judgements
Given these features, predicting car seizure

Create additional model(s) for red light ticket data

Future Work

Use cloud computing to handle full dataset

Continue tuning hyperparameters to improve the model performance

Rather than paid or not paid, predict:

- Given these features, predicting if ticket goes to court
- Given ticket goes to court court cases, predicting judgements
- Given these features, predicting car seizure

Create additional model(s) for red light ticket data

Create an applet to allow for ticket info to be entered and output a prediction

Thank you!

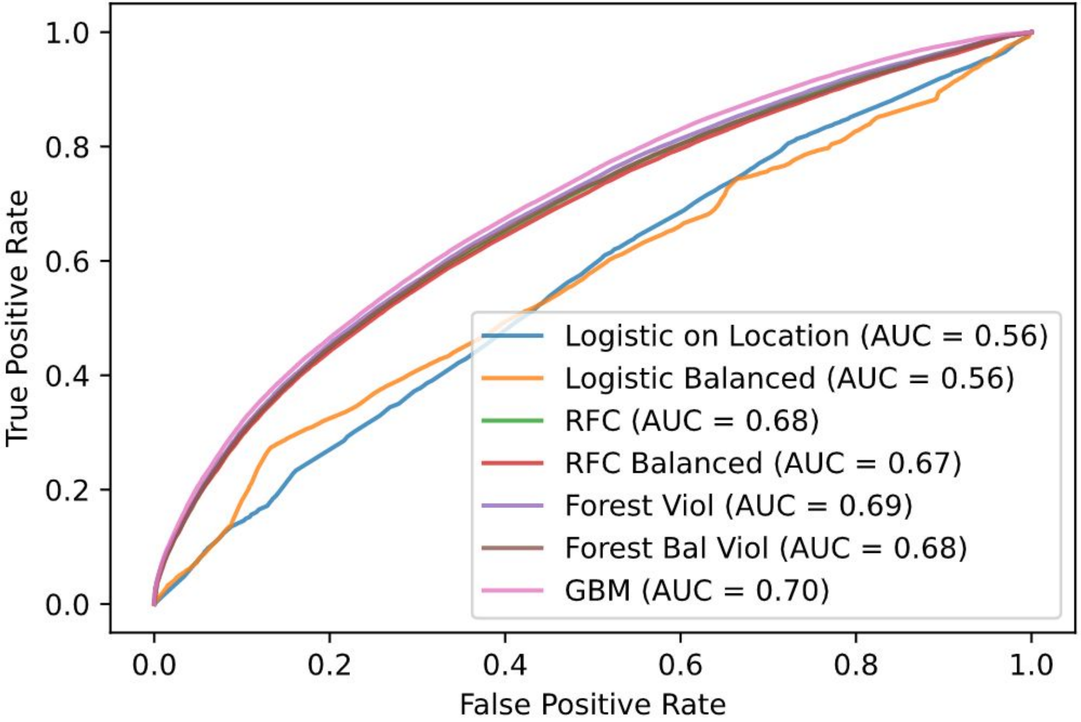
Questions?

Appendix

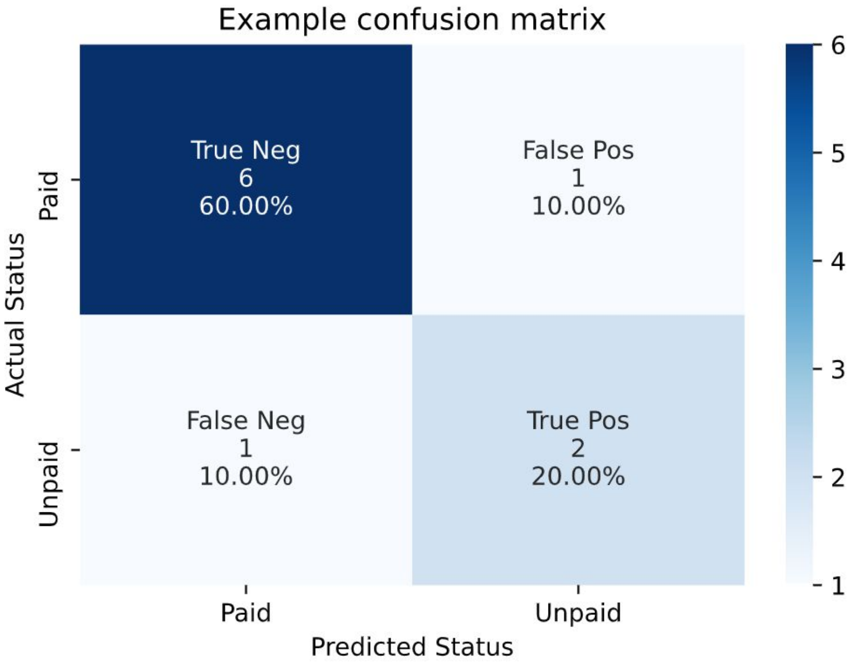
Potential Features

Ticket Number	Issue Date	Norm Address
Violation Location	License Plate Code	Year
License Plate State	License Plate Type	Month
Zip code	Violation Code	Hour
Violation Description	Unit	Warm
Unit Description	Vehicle Make	Tract ID
Fine Level 1	Fine Level 2	Community Area #
Current Amount Due	Total Payments	Community Area Name
Ticket Queue (Status)	Ticket Queue Date	Geocoded Address
Notice Level	Notice Number	Geocode Latitude
Hearing Disposition	Officer	Geocode Longitude

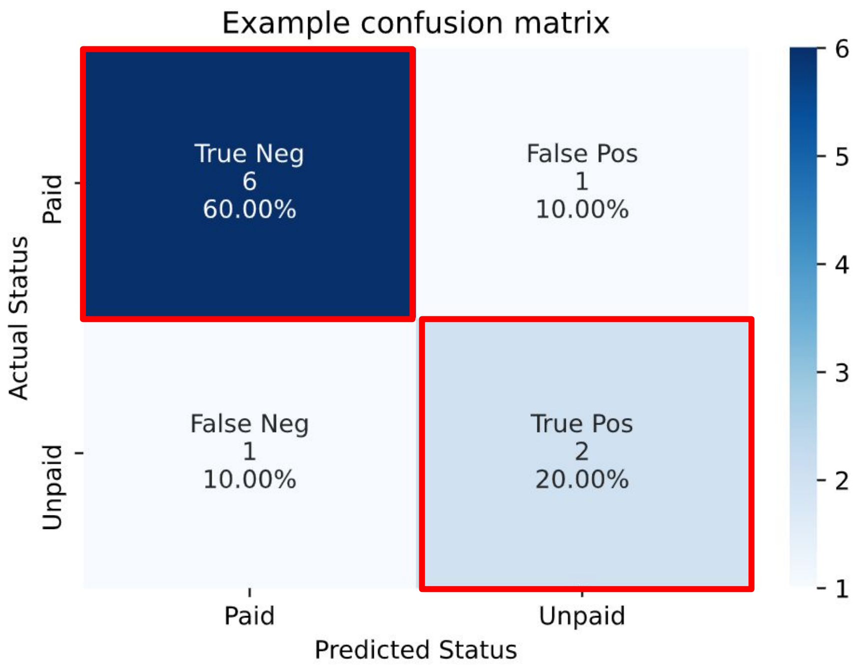
ROC Comparison (On Validation Data)



Confusion Matrix Explanation

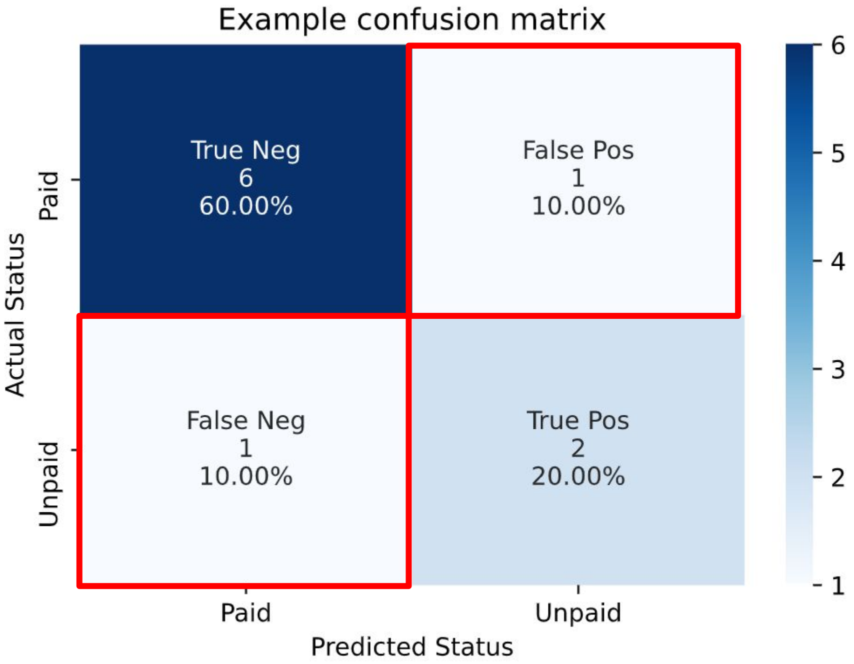


Confusion Matrix Explanation



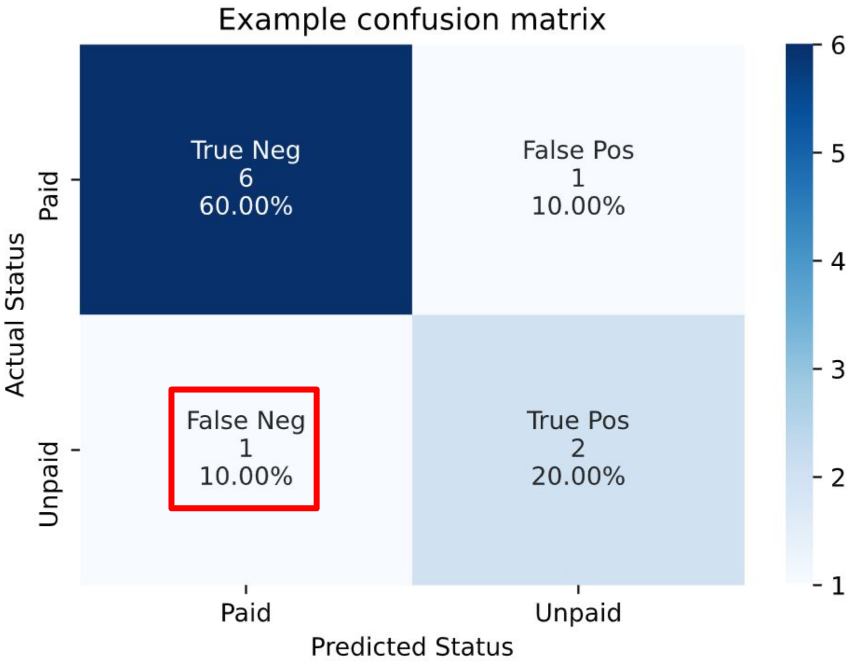
- Correct predictions on the main diagonal

Confusion Matrix Explanation



- Correct predictions on the main diagonal
- False Positives and Negatives shown on the other diagonal

Confusion Matrix Explanation



- Correct predictions on the main diagonal
- False Positives and Negatives shown on the other diagonal
- Displays number classified and percentage of the total

Error Analysis Example

Average Value Predicted Paid:	$42.89
Average Value Predicted Unpaid:	$51.60

Average Value of False Paid:	$44.87
Average Value of False Unpaid:	$46.77
Average Value of True Paid:	$45.31
Average Value of True Unpaid:	$49.80