

# NYC TAXI CLASSIFICATION

Michael Harnet

6/11/21

Metis project 4

# DATA



## January 2019

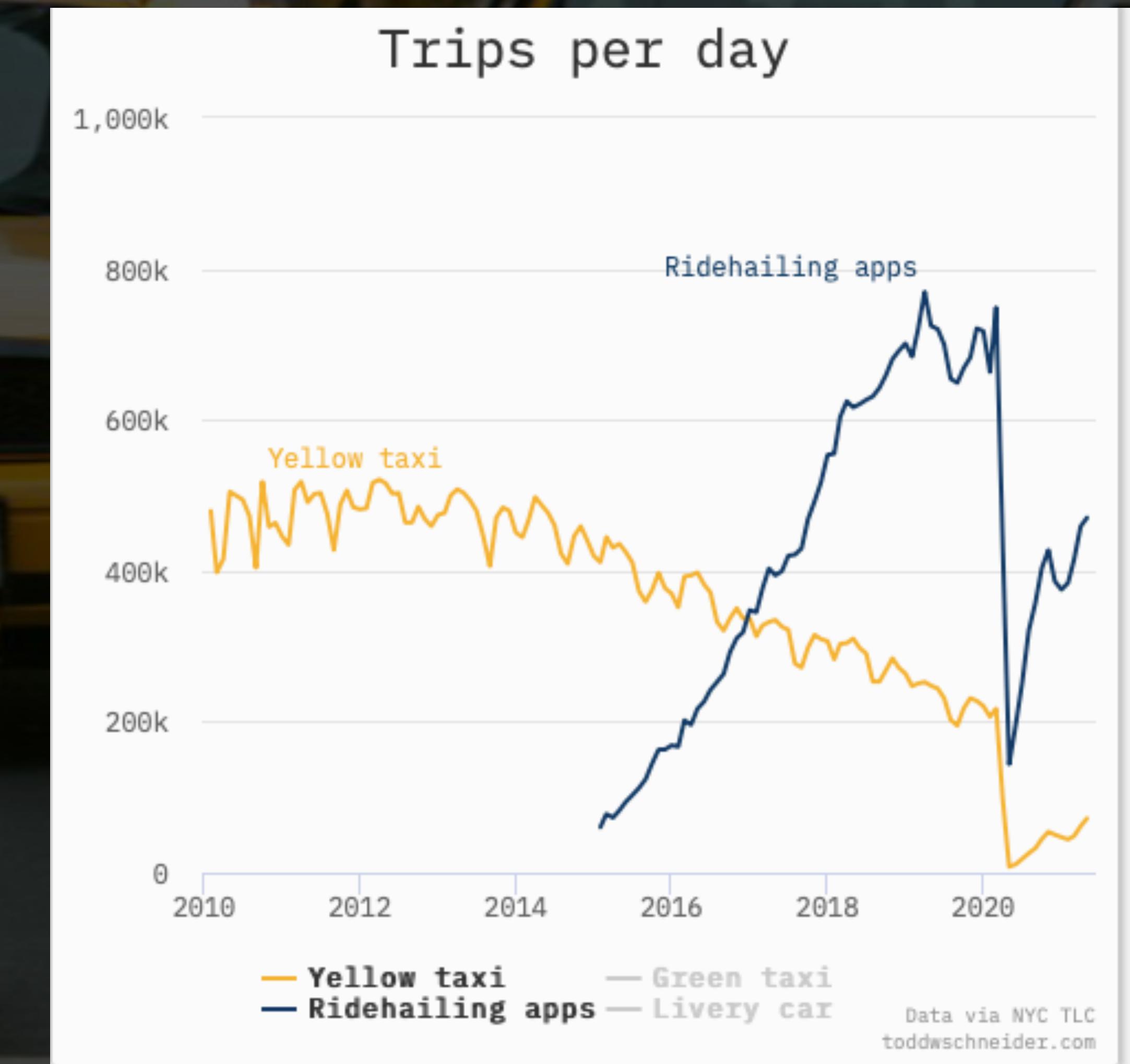
- Date and time of both pickup and drop-off
- Type of ride
- Number of passengers
- Total distance
- Total amount
- Extra charges applied
- Pickup and drop-off location

Final Dataset:

- 5.9 million observations
- 86 features

# NEED | Maximizing Profits

- Over \$100 million in revenue between yellow and green cars in January 2019 alone.
- With the introduction of ride sharing cars into the market, competition has never been stronger
- Focus on precision, minimizing false positives
- Graph pulled from [toddwschneider.com](http://toddwschneider.com)



# Initial Scoring

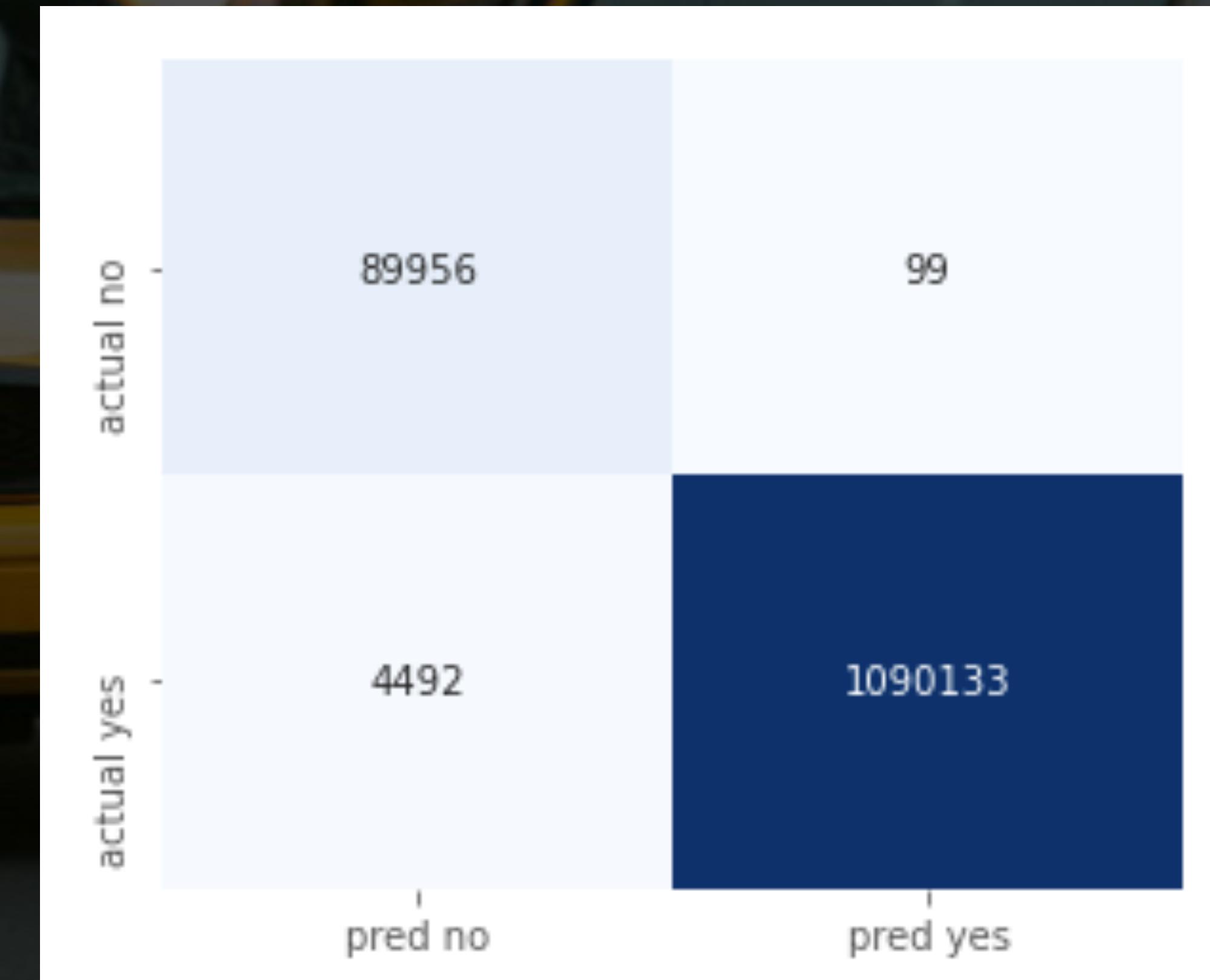
- Metrics across all baseline models were performing very well.
- Still too many false positives
- Missing categorical features for interpretation

# Logistic Regression Confusion Matrix

		No Tip	Tip
No Tip	166273		1114
	1074	393902	
No Tip		1114	1074

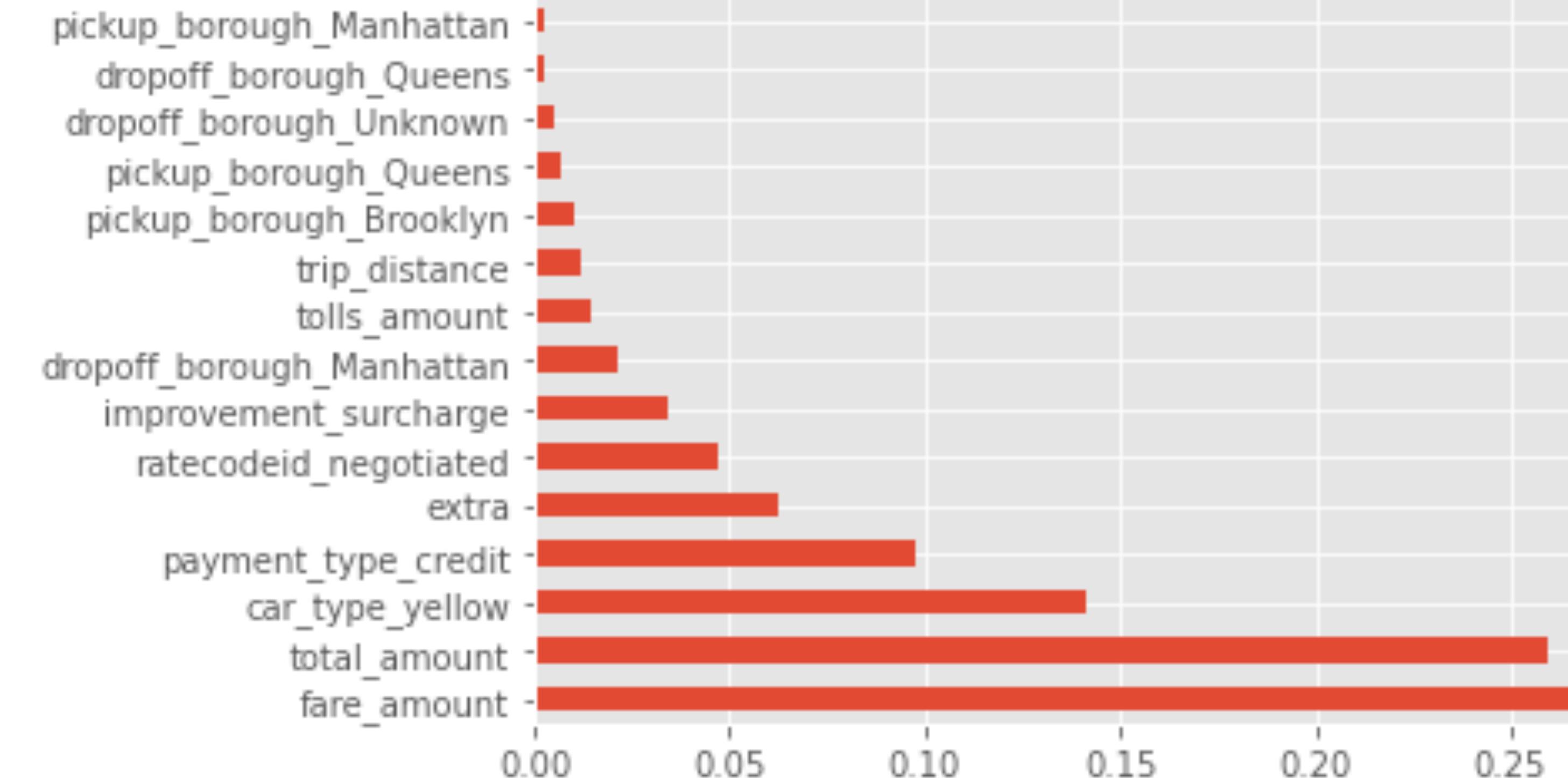
# Continuing with Logistic Regression

- After adding all categorical features, false positives jumped significantly to over 8 thousand.
- After tuning Hyperparameters, I was able to drop the false positives to under
- Precision and Recall stayed at 99%
- Log loss was 2.4



# Feature importance

- Using the tree models I also worked with, I extracted feature importances
- Most impactful features include
  - Fare amount
  - Being a yellow cab
  - Paying with credit
  - Negotiating your own fare
  - Being dropped off in Manhattan



# Future work

- In the future I would like to move to a cloud computing service, to allow these models to be done on a larger scale
- Combining multiple models together into an ensemble to see if it improves scores.



Thank You