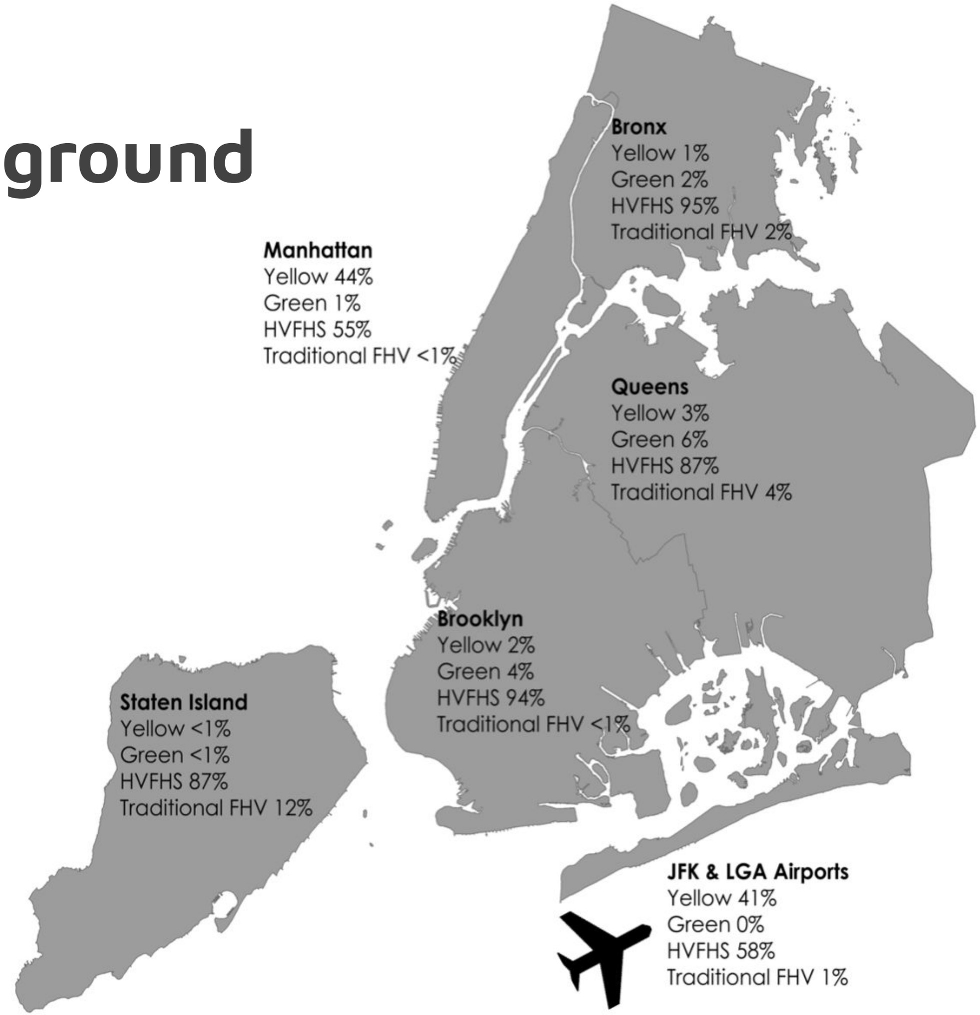
STAT 453: NYC Taxi Tip Prediction

Clarity Kummer, Bruce Peng, Abigail Sikora, Angela Wei

Background

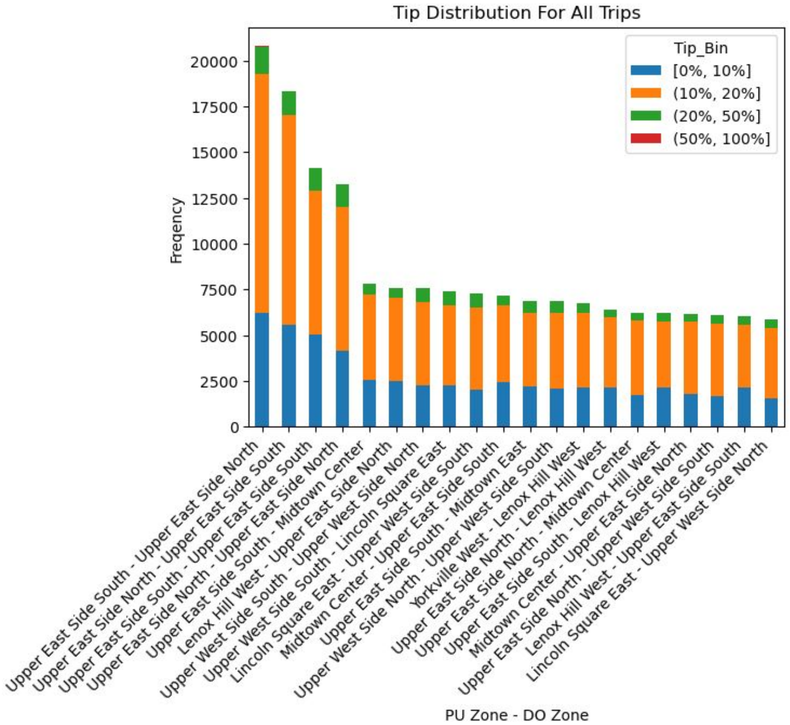

Bronx
Yellow 1%
Green 2%
HVFHS 95%
Traditional FHV 2%
Manhattan
Yellow 44%
Green 1%
HVFHS 55%
Traditional FHV <1%
Queens
Yellow 3%
Green 6%
HVFHS 87%
Traditional FHV 4%
Brooklyn
Yellow 2%
Green 4%
HVFHS 94%
Traditional FHV <1%
Staten Island
Yellow <1%
Green <1%
HVFHS 87%
Traditional FHV 12%
JFK & LGA Airports
Yellow 41%
Green 0%
HVFHS 58%
Traditional FHV 1%

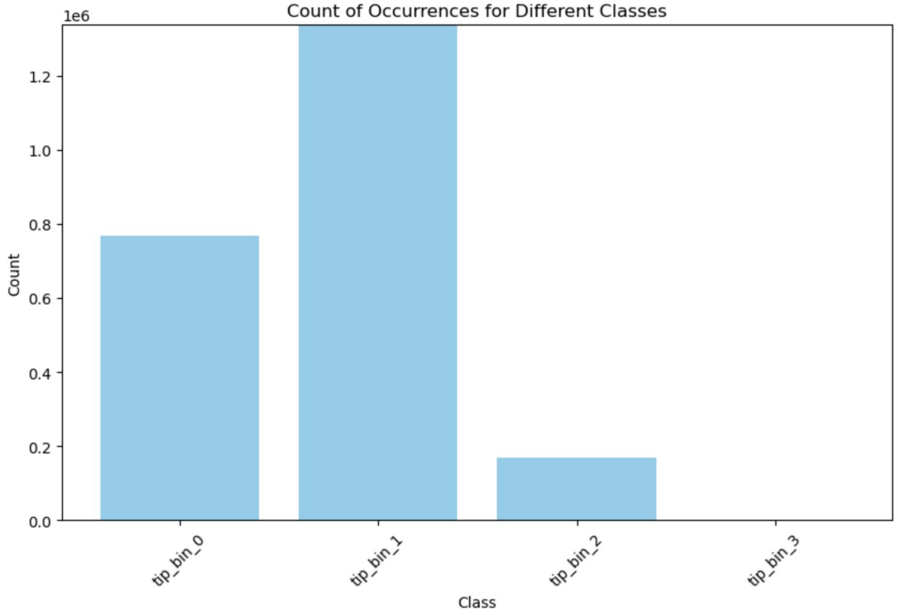
Can we predict how well people tip for yellow taxi trips in Manhattan in June of 2021?

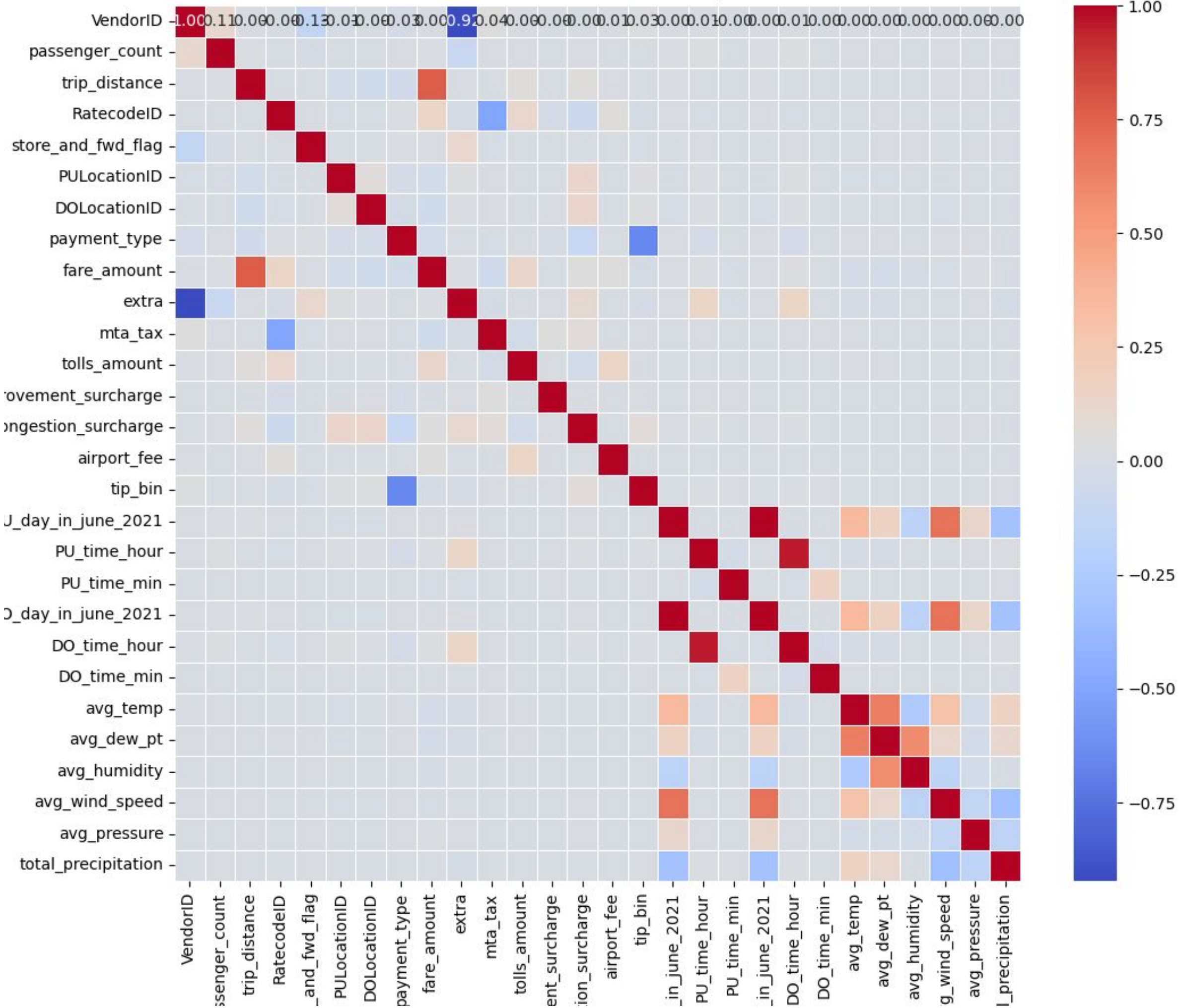
What does the data look like?

- **Taxi Info:** Vendor, Rate Code ID, Store And Forward Trip
- **Passenger Trip Info**: Pickup Date/Time, Pickup Location, Drop-off Date/Time, Drop-off Location, Trip Distance, Passenger Count
- **Payment Info**: Payment Type, Fare Amount, Extra Cost, MTA Tax, Tolls Amount, Improvement Surcharge, Congestion Surcharge, Airport Fee, Tip Range
- **Daily Average Weather**: Temperature, Dew Point, Humidity, Wind Speed, Pressure, and Total Precipitation

Exploratory Data Analysis







Correlation Matrix Heatmap

List of Models

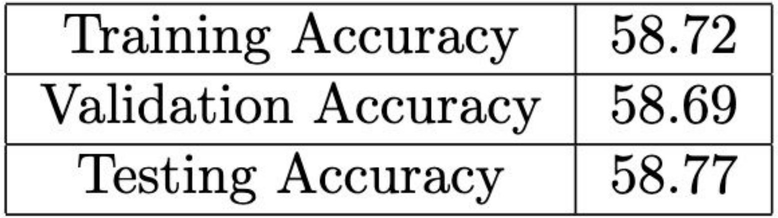
Multilayer Perceptron Neural Network

Decision Tree Classifier/ Random Forest

Softmax

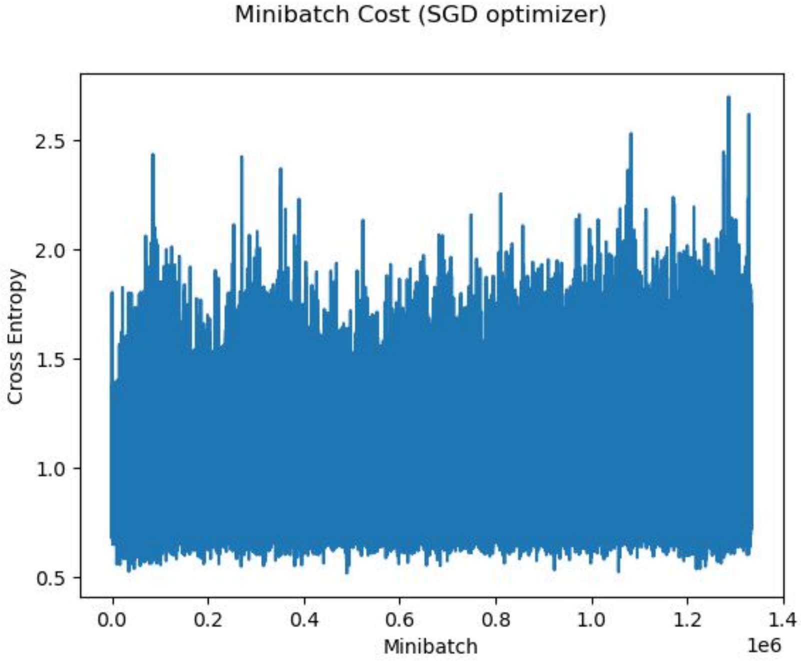
Multinomial Logistic Regression

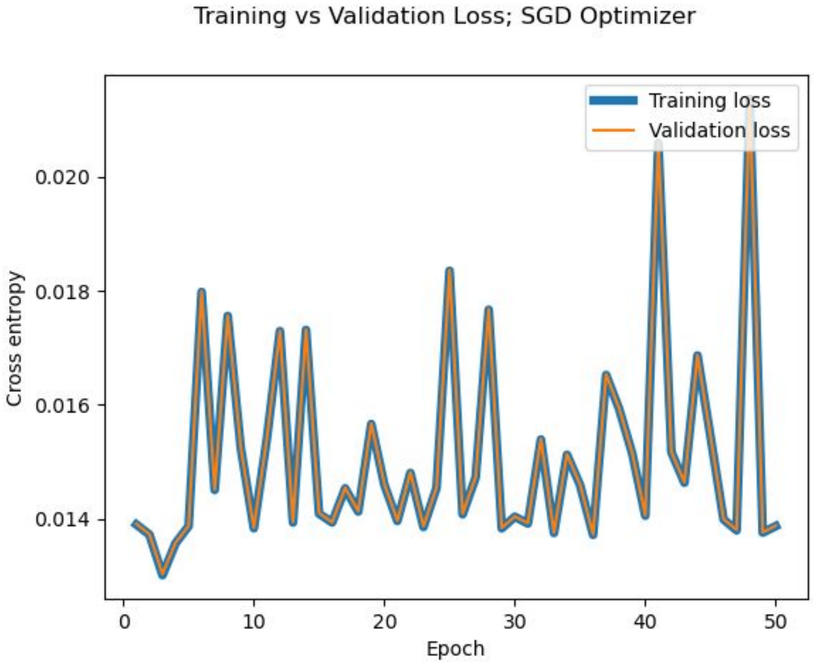
Neural Network



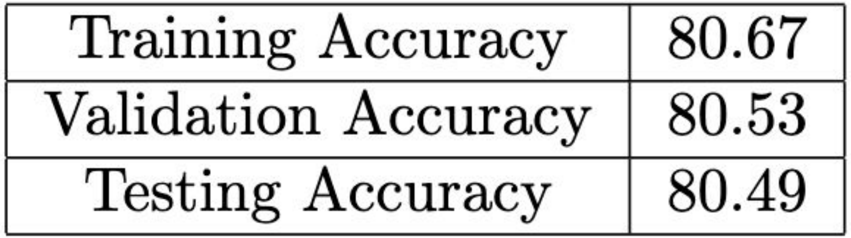
Training Accuracy	58.72
Validation Accuracy	58.69
Testing Accuracy	58.77

Table 1: Optimizer: SGD



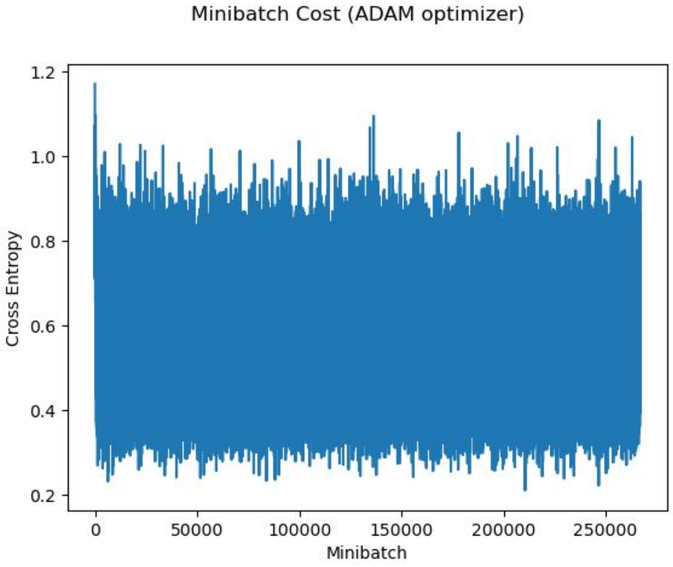


Neural Network

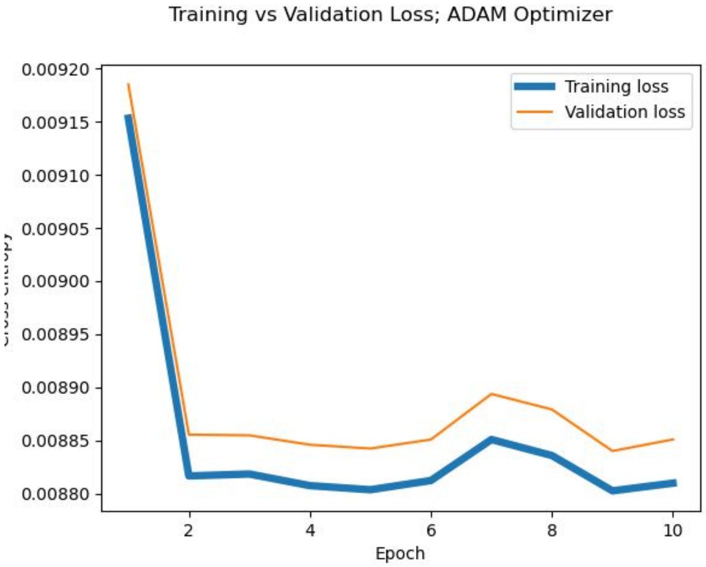


Training Accuracy	80.67
Validation Accuracy	80.53
Testing Accuracy	80.49

Table 1: Optimizer: ADAM

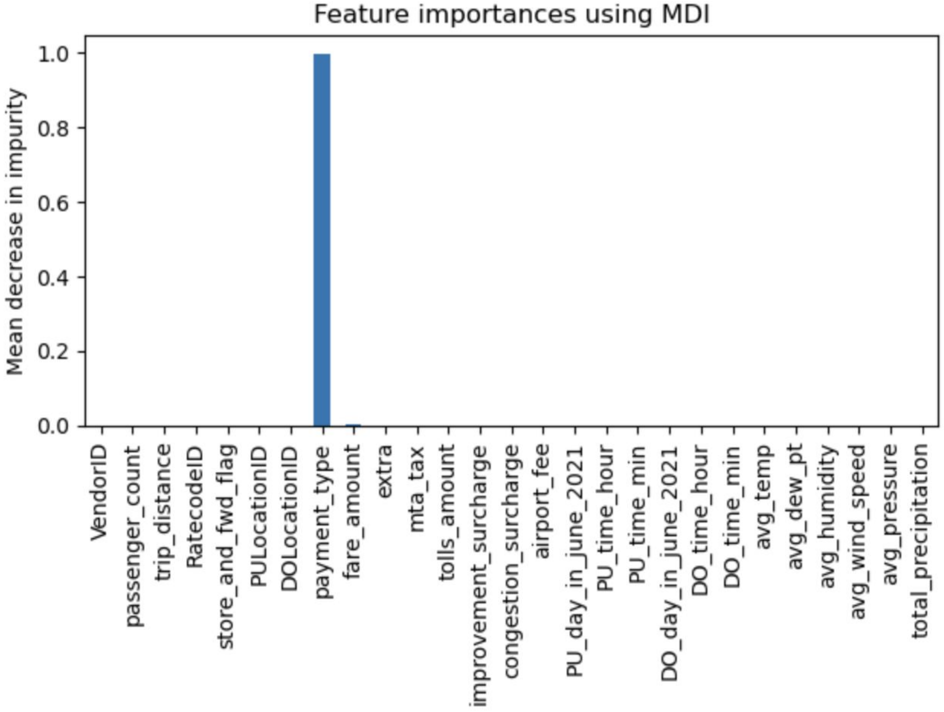


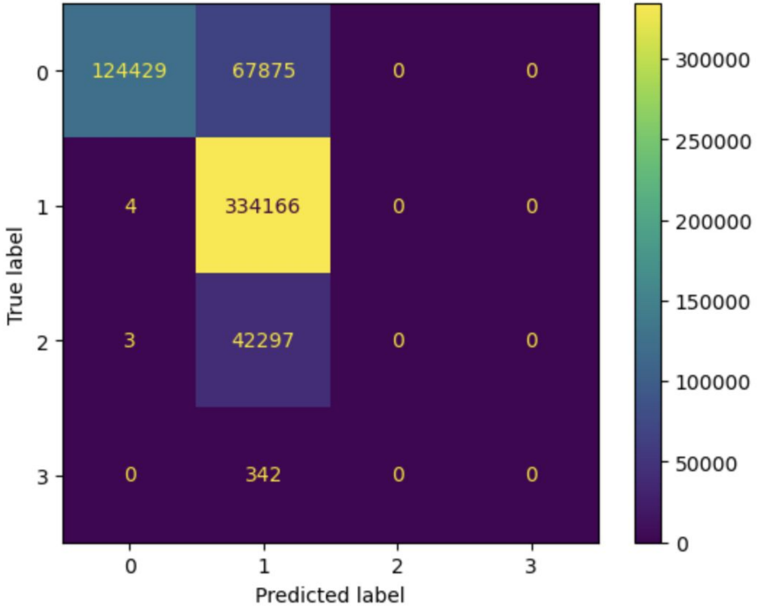
Minibatch Cost (ADAM optimizer)



Training vs Validation Loss; ADAM Optimizer

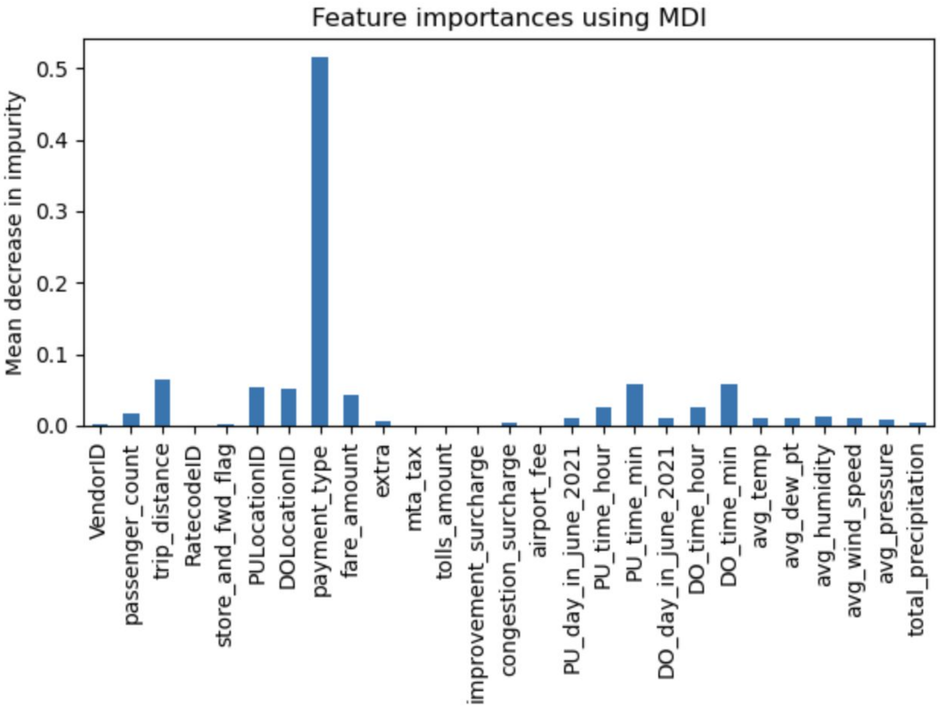
Decision Tree (Test R^2 = 0.8058)

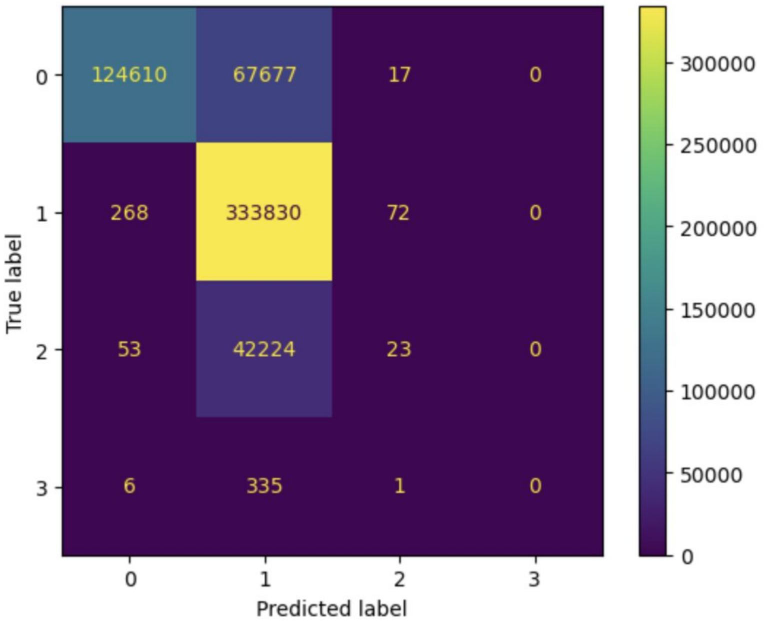




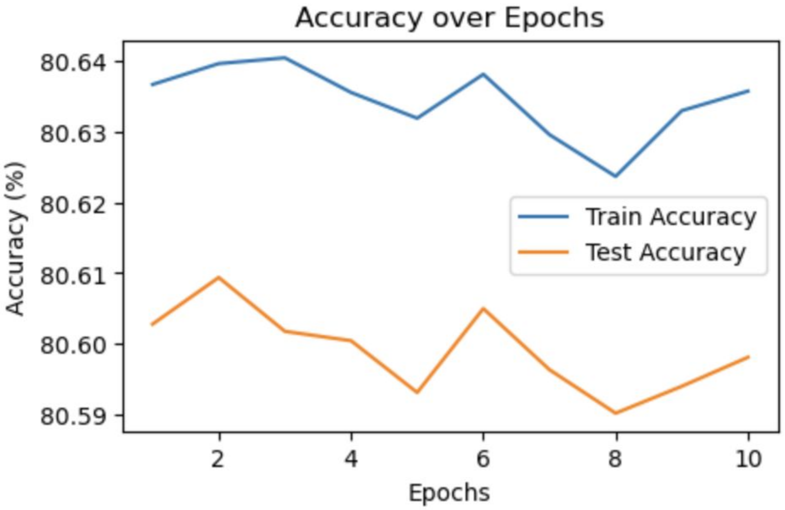
True label \ Predicted label	0	1	2	3
0	124429	67875	0	0
1	4	334166	0	0
2	3	42297	0	0
3	0	342	0	0

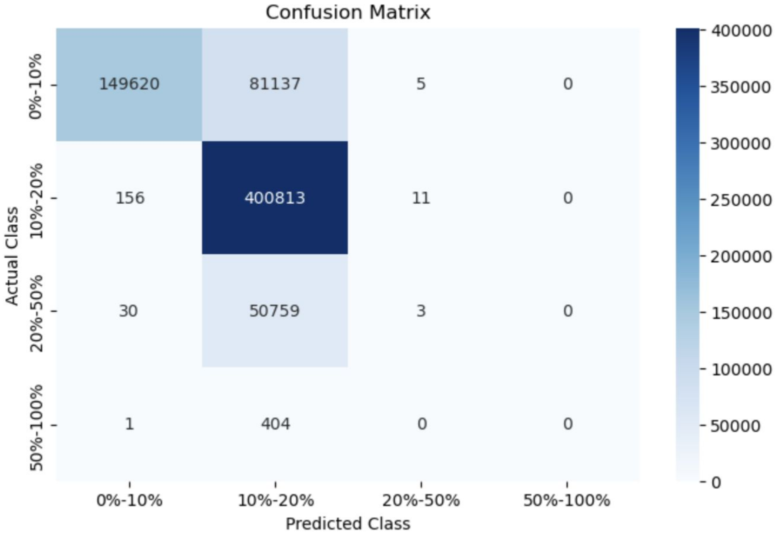
Random Forest (Test R^2 = 0.8056)





Softmax Regression


Accuracy over Epochs

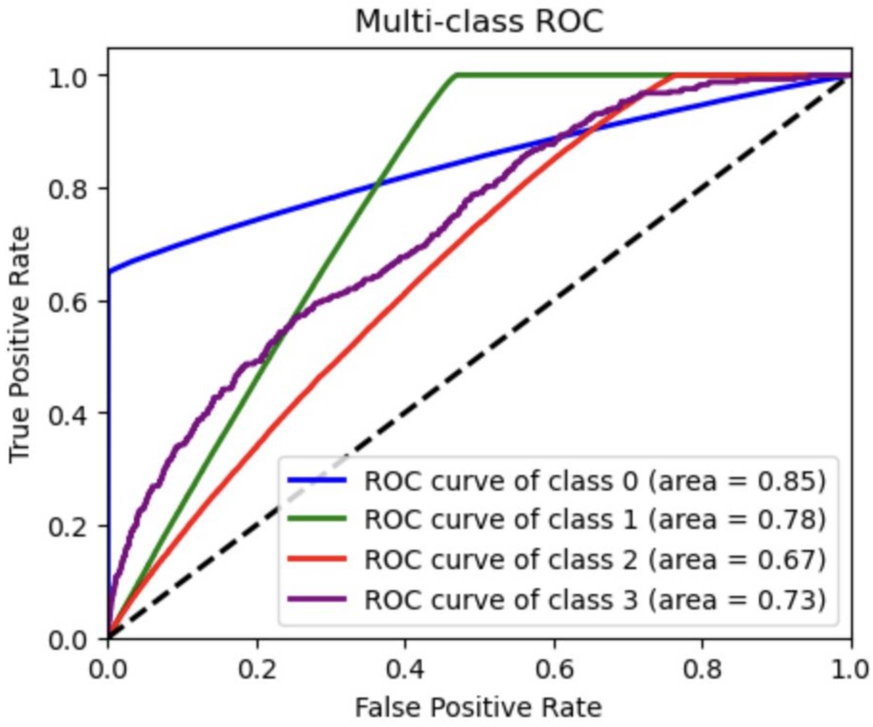

Confusion Matrix

Actual Class \ Predicted Class	0%-10%	10%-20%	20%-50%	50%-100%
0%-10%	149620	81137	5	0
10%-20%	156	400813	11	0
20%-50%	30	50759	3	0
50%-100%	1	404	0	0

Softmax Regression

Advanced Method(**Receiver Operating Characteristic, ROC):**

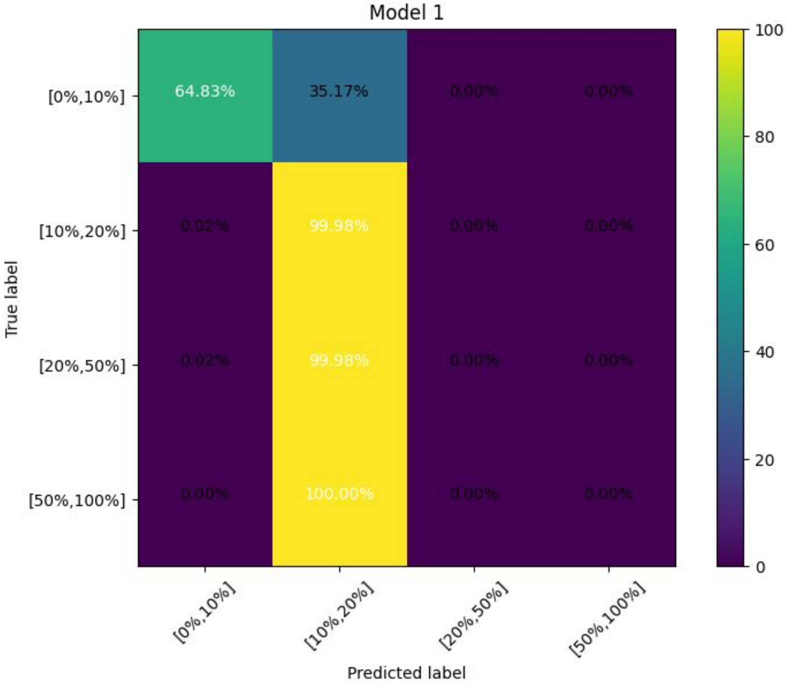
Definition: The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.



Multinomial Logistic Regression

Model Identifier	Solver	Penalty	Training Time	Accuracy
Model 1	lbfgs	l2	~6 mins	80.61%
Model 2	lbfgs	None	~9 mins	80.61%
Model 3	newton-cg	None	~18 mins	80.61%
Model 4	newton-cg	l2	~16 mins	80.61%
Model 5	sag	None	~85 mins	80.61%
Model 6	saga	None	~93 mins	80.61%

Multinomial Logistic Regression Confusion Matrix for Model 1



Conclusion/ Recommendations

Traditional ML works well for tabular data (NN not as much)

- Apply the models to more data (i.e. for hire trip reviews, driver ratings, etc.)
- Do more models (TabPFN, TP-BERTa, etc.)
- Group tips into more buckets

Thank you!