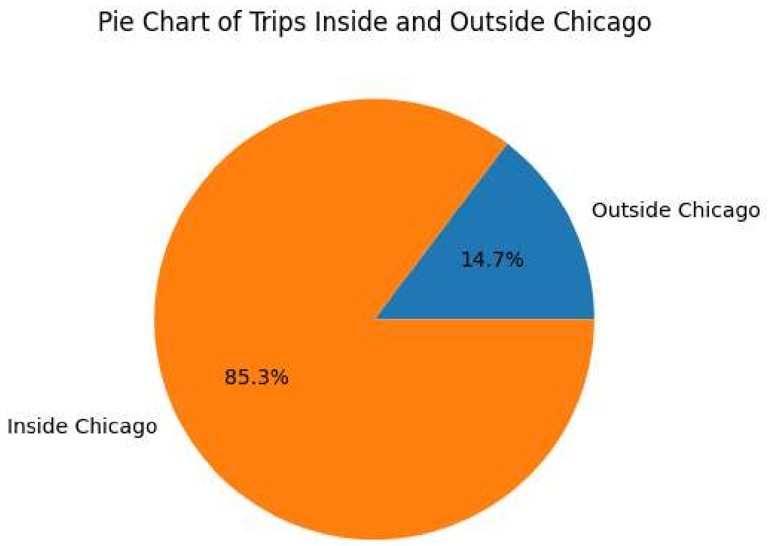
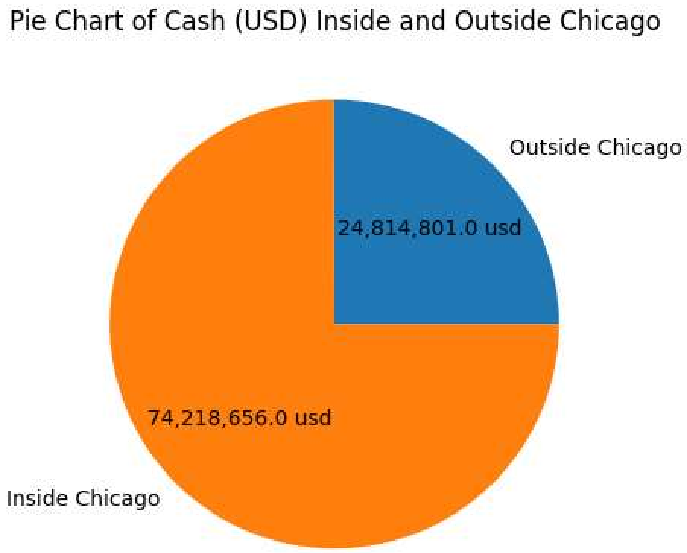
Predicting Fares

The opportunity

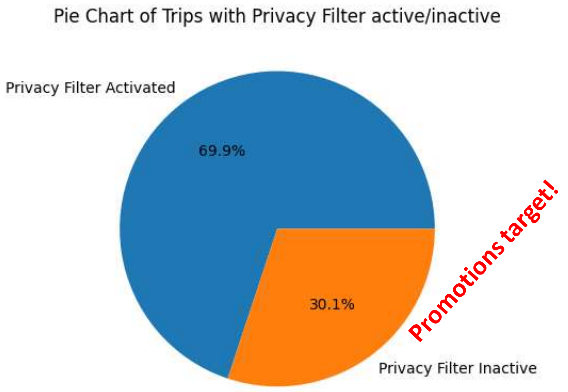
- 85% of rides are inside Chicago, which amounts to a net income of roughly 74.2 million USD

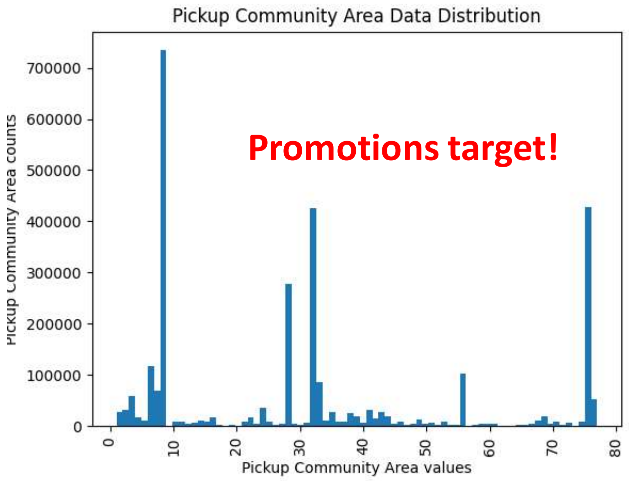


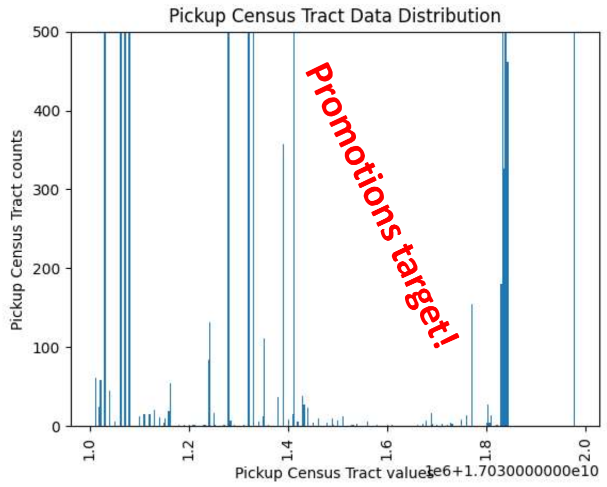


The value of the data collected (1/2)

- More than 1 million rides do share all the information (privacy filter inactive).

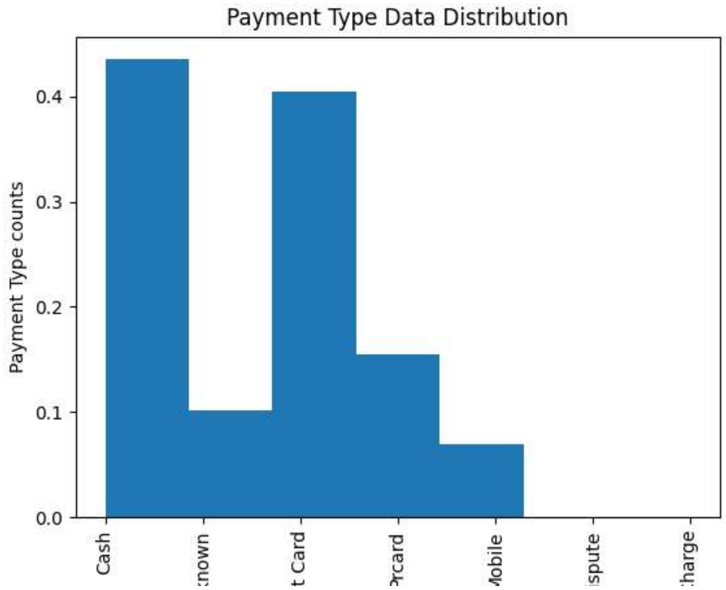


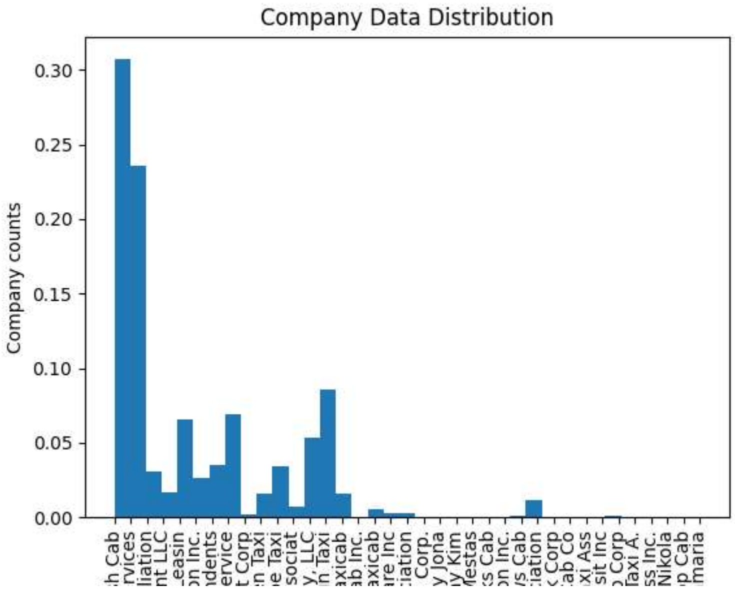




The value of the data collected (2/2)

- Partnerships and deals can be done with credit card companies which amount to almost 40% of the payment methods used. Also with Taxicab companies.

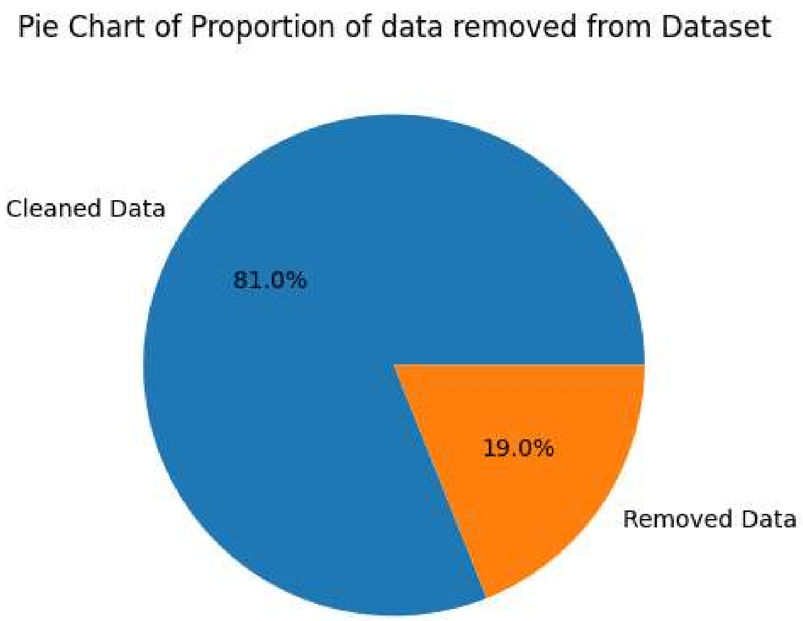




Fare Prediction: Data preprocessing

- Removing missing fares (NaN/Null).
- Removing missing fields (NaN/Null).
- Removing outliers (IQR).
- Fixing fields (values=0)
- Removing unnecessary fields.

	Number of samples
Missing Fares	608
Missing Fields	117704
Hiden (Privacy)	2759599
Complete Samples (all fields)	1070134
Total	3948045

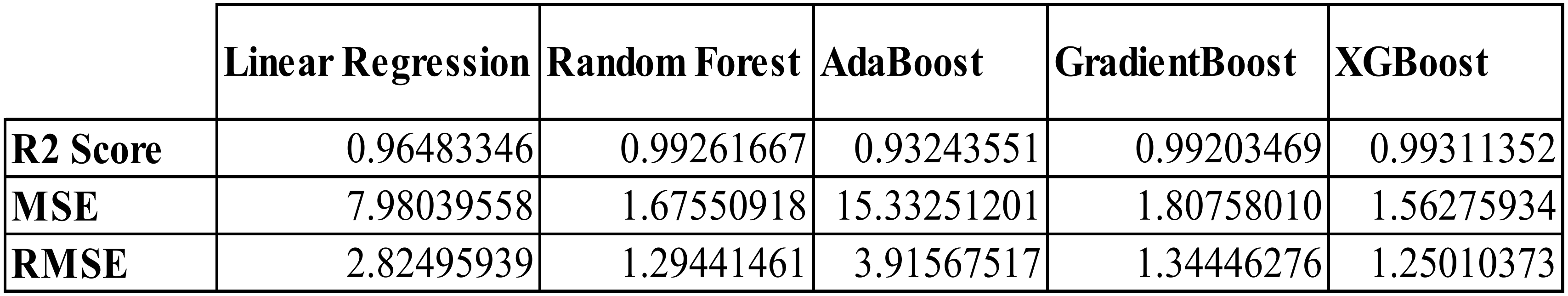


Fare Prediction: Data preparation

- 2 datasets were prepared:
  - **Inside Chicago dataset**: containing only samples inside Chicago, 851,154 samples (Table 4).
  - **Whole dataset**: all the samples combined inside/outside Chicago 3,199,179 samples (Table 4).

Fare Prediction: Benchmark models

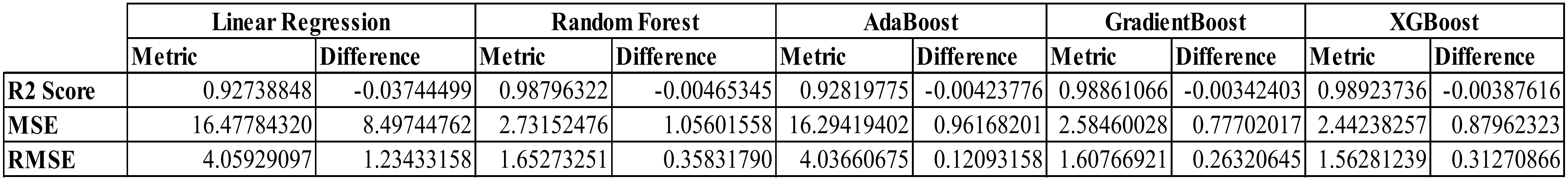
- Trained with all features/fields



	Linear Regression	Random Forest	AdaBoost	GradientBoost	XGBoost
R2 Score	0.96483346	0.99261667	0.93243551	0.99203469	0.99311352
MSE	7.98039558	1.67550918	15.33251201	1.80758010	1.56275934
RMSE	2.82495939	1.29441461	3.91567517	1.34446276	1.25010373

Fare Prediction: Removing fields

- All fields were removed except for [“Trip Miles”, “Trip Seconds”, “Company”, “Payment Type”]. The comparison with the retrained models is:



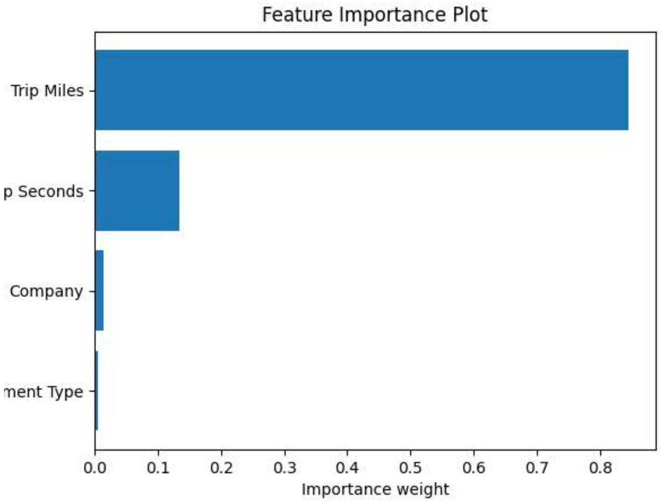
	Linear Regression		Random Forest		AdaBoost		GradientBoost		XGBoost	
	Metric	Difference	Metric	Difference	Metric	Difference	Metric	Difference	Metric	Difference
R2 Score	0.92738848	-0.03744499	0.98796322	-0.00465345	0.92819775	-0.00423776	0.98861066	-0.00342403	0.98923736	-0.00387616
MSE	16.47784320	8.49744762	2.73152476	1.05601558	16.29419402	0.96168201	2.58460028	0.77702017	2.44238257	0.87962323
RMSE	4.05929097	1.23433158	1.65273251	0.35831790	4.03660675	0.12093158	1.60766921	0.26320645	1.56281239	0.31270866

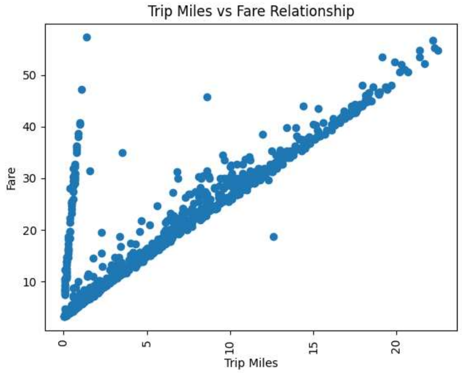
- **Conclusion**: : in this analysis it was shown that most features in the dataset are not relevant in the estimation of the Fare. *In the estimation of the Fare less than 0.5% accuracy loss is observed compared with benchmark models in the previous slide.*

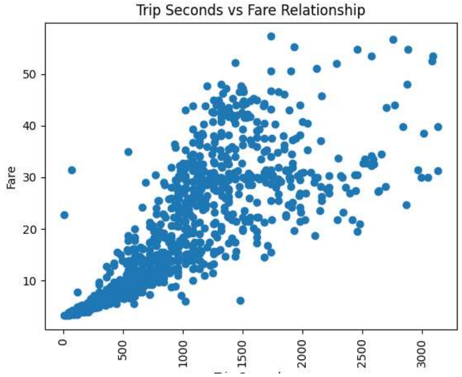
Fare Prediction: Final Model

- XGBoost with all the data available but only considering [“Trip Miles”, “Trip Seconds”, “Company”, “Payment Type”].

	XGBoost (Pooled Data)
R2 Score	0.97123309
MSE	5.47489559
RMSE	2.33984948


Feature Importance Plot


Trip Miles vs Fare Relationship


Trip Seconds vs Fare Relationship