Launching a Ride-Hailing Service in Chicago

The opportunity!

The data is composed by trips inside as well as outside of Chicago for a period of 1 year (2021). To get a good grasp of the "opportunity", a pie chart is provided in Figure 1 (left) where we can appreciate that more than 85% of the data is for trips inside of Chicago (roughly 3.3 million rides). The net cash coming from the trips inside and outside of Chicago amounts to *~74.2 million USD*, and *~24.8 million USD* respectively (Figure 1 right).

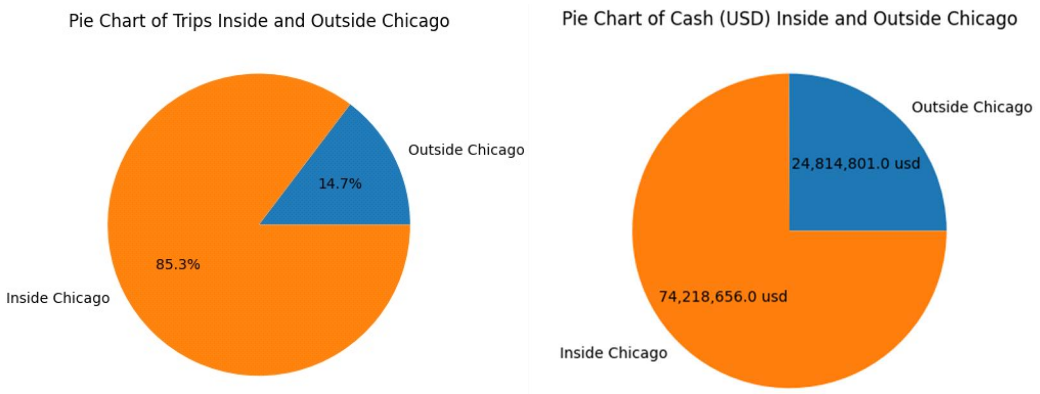

Figure 1. Trips proportions inside/outside Chicago.

In terms of fares, it is worth noting that fares inside Chicago tend to be lower compare to those outside. This can be clearly appreciated in the boxplot in Figure 2. However, further analysis to understand whether the fares do depend on rides inside/outside Chicago is required.

Figure 2. Fare boxplot for trips inside/outside Chicago

The value of the data collected

In terms of data collection for each ride, there is a total of 23 fields that could be registered per ride. However, due to privacy and/or error in the system, it is likely that we are going to have incomplete data points, where some fields are going to be missing. To understand how much of a problem this is, please

refer to the Table 1. If we combine Inside/Outside Chicago data (Figure 3) we can clearly appreciate that around ~69.9% of the customers prefer to enable privacy, this will avoid collecting certain fields such as [“Pickup Census Tract”, “Dropoff Census Tract”], as for missing/incomplete samples represent only ~2.9% of the whole dataset.

Table 1. Proportion of data Hidden, Incomplete, and Complete in the dataset

	Inside Chicago	Outside Chicago
Hidden (privacy)	2297312	462287
Incomplete	123	117581
Complete	1070134	0
Total	3367569	579868

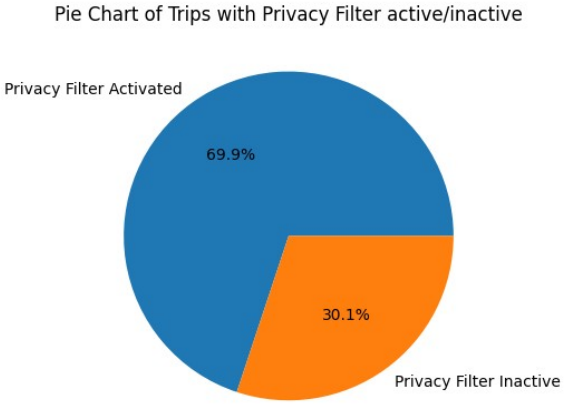

Figure 3. Proportion of data that is hidden because of privacy.

The information shown in the previous paragraph is of great importance if we would like to also consider targeted promotion campaigns for areas where [“Pickup Census Tract”, “Dropoff Census Tract”] could be available motivating potential customers in the region to use the service. The distribution per census tract is shown in Figure 4 (limited to 500 data points per census tract or visualization purposes). Similarly, can be done for fields such as [“Pickup Community Area”, “Dropoff Community Area”] where the distribution is also illustrated in Figure 5.

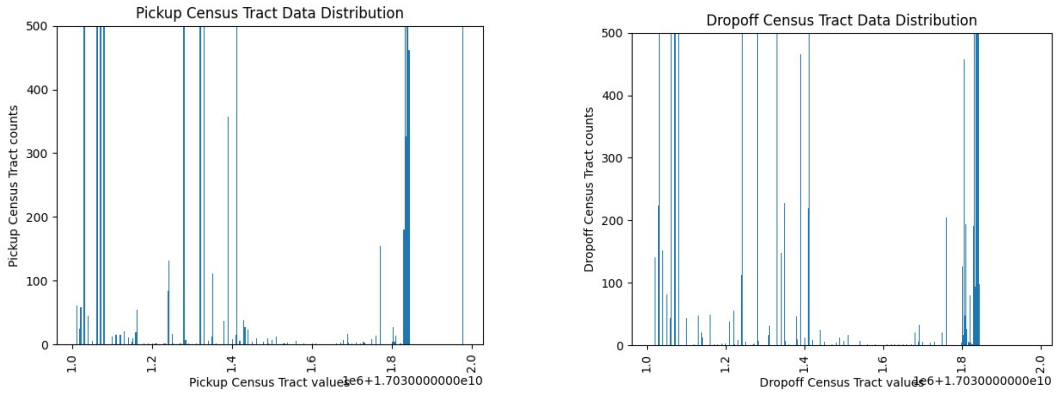

Figure 4. Data distribution per Census Track (pickup, dropoff)

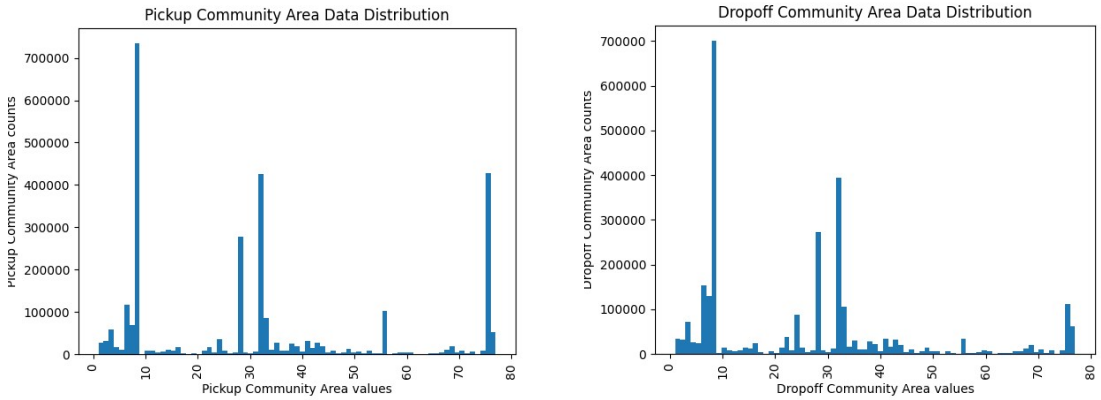

Figure 5. Data Distribution per Community Area (pickup, dropoff)

Additionally, specific deals/promotions can be designed with credit cards (~40% of total payments) companies and taxicab companies which serve most of the ride (Figure 6). Also incentives can be offered to promote taxicab companies with small volumes of rides.t

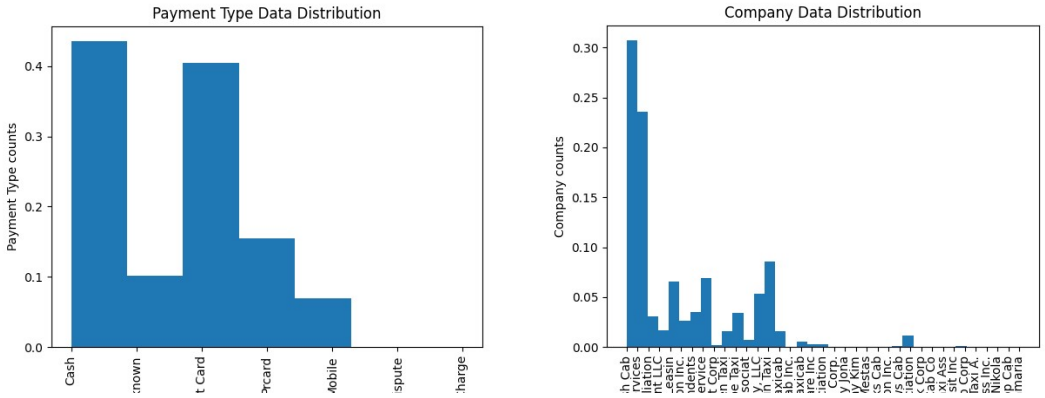

Figure 6. Data distribution for payment type (left) and taxcab company (right).

The service

Finally, a very important component of the service is to determine the "Fare" for each trip as accurately as possible. This is a very important element for user satisfaction; therefore, to determine the "Fare" is it essential to preprocess the data and determine how much information each field (feature) provide in determining "Fare" for the trip. This is going to be detailed in the following section of this report.

But, before moving to the next section, it is important to mention that other field can be used to improve not only customer satisfaction, but also, driver satisfaction; but this is a topic I will not cover in this report.

Developing a Model to Predict Fares

1. **Data preprocessing**
    a. **Removing missing "Fares"**: since we are predicting "Fares" it is important to remove data points where this value is not available. The reason being, this samples cannot be used for training any model. Here it is important to mention that whether the data point is inside/outside Chicago is not relevant. The amount of samples removed was 608, which only represents 0.01% of the data available (3,948,045), please refer to Table 2.
    b. **Removing missing fields**: here we need to be careful since not all missing fields mean "missing samples", due to privacy several fields in the dataset will be effectively missing. After careful filtering I was able to identify 117,704 samples where fields such as ["Pickup/Dropoff Census Track", "Pickup/Dropoff Community Area"] are missing for samples where no privacy is activated. This represents just 2.9% of the total data available (Table 2).
    c. **Outlier removal**: for fields such as ["Trip Miles", "Trip Seconds", "Fare"] will contain values for which the probability of occurrence is just too low. To make sure we don't have these kind of samples in the dataset, I will use Interquartile Range (IQR) to perform removal of outliers where *LowerBound = max(0, Q1-1.5*IQR)* and *UpperBound = Q3+1.5*IQR*. For reference on the values please check Table 3. Additionally, if ["Total Trip", "Taxi ID"] fields are missing I will also remove this samples from the dataset. Finally, the amount of data points removed was 748,866 samples which represents around 19% of the data available (Figure 7).
    d. **Fixing fields**: for fields such as ["Tips", "Tolls", "Extras"] can have missing values, and does not mean these samples are not usable. On the contrary, it requires fixing, which in my case I will assign "0".
    e. **Removing unnecessary fields**: Some fields are likely to provide no information at all in the estimation of the fare, these fields/columns are to be removed, and they are specified as follow
        i. "**Pickup Centroid Latitude**": Trip Miles is likely to have more accurate information.
        ii. "**Pickup Centroid Longitude**": Trip Miles is likely to have more accurate information.
        iii. "**Dropoff Centroid Latitude**": Trip Miles is likely to have more accurate information.
        iv. "**Dropoff Centroid Longitude**": Trip Miles is likely to have more accurate information.
        v. "**Pickup Centroid Location**": Latitude and Longitude provide the same information
        vi. "**Dropoff Centroid Location**": Latitude and Longitude provide the same information
        vii. "**Time Start Timestamp**": Trip Seconds provides more relevant information
        viii. "**Time End Timestamp**": Trip Seconds provides more relevant information
        ix. "**Trip Total**": This is a combination of "Fare" + "Tips" + "Tolls" + "Extras" therefore we should not use this otherwise the model will give all the weight to this feature.

f. **Dataset preparation**: As I don't know what is the effect of data inside/outside Chicago in determining the "Fare", I create 2 datasets. The logic behind it is to check what is the effect of the different features/fields of the data when predicting "Fare" with different Machine Learning models. If the effects of fields critical in determining if the samples as inside/outside Chicago is very small, we can confidently remove these fields and use all the data available.
    i. **Inside Chicago dataset**: containing only samples inside Chicago, 851,154 samples (Table 4).
    ii. **Whole dataset**: all the samples combined inside/outside Chicago 3,199,179 samples (Table 4).

Table 2. Data removal for samples with no "Fare" or missing field when privacy is not activated

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

Table 3. IQR outlier removal for ["Trip Miles", "Trip Seconds", "Fare"]

	IQR	LoweBound	UpperBound
Fare	23.25	0	65.375
Trip Miles	9	0	23.2
Trip Seconds	1130	0	3315

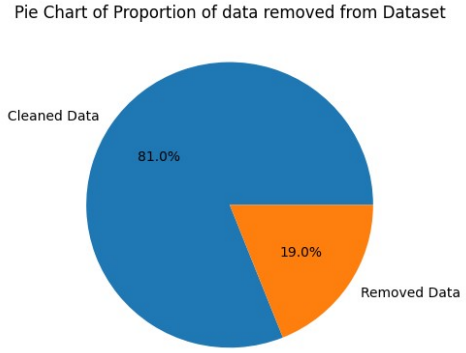

Figure 7. Data removed through IQR and missing ["Total Trip", "Taxi ID"] fields

Table 4. Samples available for model training/testing

	Number of samples
Inside Chicago	851154
Whole Dataset	3199179

2. **Features Analysis through different Models**
    a. **Models tested**: To determine how relevant/important specific features/fields in the dataset are, I will test different model, where I will train the model with/without the feature to determine how much the accuracy of the model will drop. To measure this accuracy, I will use R2 score which describe the amount of variance in the dependent variable (Fare) that is predictable from the independent variables (Features/Fields). Additionally, the MSE and RMSE are going to be shown. The list of models to be tested are:
        i. **Linear Regression**: as this is a regression problem.
        ii. **Random Forest (bagging)**: as this is a regression problem.
        iii. **AdaBoost**: as this is a regression problem.
        iv. **GradientBoost**: as this is a regression problem.
        v. **XGBoost**: as this is a regression problem.
    b. **Features/Fields considered/removed**: the fields are
        i. *'Trip ID'*: to be removed always as it is not necessary.
        ii. *'Taxi ID'*: since information such as car type may be intrinsically encoded, it makes sense to consider it.
        iii. *'Trip Seconds'*: very likely to be important in the estimation of the Fare.
        iv. *'Trip Miles'*: very likely to be important in the estimation of the Fare.
        v. *'Pickup Census Tract'*: I don't know if privacy affects the estimation of the Fare.
        vi. *'Dropoff Census Tract'*: I don't know if privacy affects the estimation of the Fare.
        vii. *'Pickup Community Area'*: I don't know if trips inside/outside Chicago do impact the estimation of the Fare.
        viii. *'Dropoff Community Area'*: I don't know if trips inside/outside Chicago do impact the estimation of the Fare.
        ix. *'Tips'*: I don't know if tips are relevant in the estimation of the Fare, logic says no.
        x. *'Tolls'*: I don't know if tips are relevant in the estimation of the Fare, logic says no.
        xi. *'Extras'*: I don't know if tips are relevant in the estimation of the Fare, logic says no.
        xii. *'Payment Type'*: I don't know if tips are relevant in the estimation of the Fare, logic says very likely.
        xiii. *'Company'*: I don't know if tips are relevant in the estimation of the Fare, logic says very likely.
    c. **Models**: the approach will be to train a benchmark model where all the features/fields are considered, and then re-train the same models without them, then compare R2 to determine if the feature contributes significantly with the estimation of the Fare.
        i. **Benchmarks**: from Table 5 it is easy to observe that Random Forest, GradientBoost, and XGBoost are the best models where ~99% of the variance from the dependent variable (Fare) can be predicted when considering all independent variables (features/fields).

Table 5. Models *benchmark* trained with all features

	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

ii. **No Taxi ID in the features**: from Table 6, in all the models it can be observed that *Taxi ID does not have significant impact in the estimation of the Fare where the difference with the benchmark models is less than 0.2% for most of the models*, correcting my initial belief where I thought the car type may be somehow encoded.

Table 6. Models retrained without Taxi ID feature

	Linear Regression		Random Forest		AdaBoost		GradientBoost		XGBoost	
	Metric	**Difference**	**Metric**	**Difference**	**Metric**	**Difference**	**Metric**	**Difference**	**Metric**	**Difference**
R2 Score	0.93787899	-0.02695448	0.99208094	-0.00053573	0.93411774	0.00168223	0.99120012	-0.00083457	0.99235914	-0.00075438
MSE	14.09721574	6.11682016	1.79708418	0.12157500	14.95076014	-0.38175187	1.99697116	0.18939106	1.73395090	0.17119156
RMSE	3.75462591	0.92966652	1.34055368	0.04613907	3.86662128	-0.04905388	1.41314230	0.06867954	1.31679569	0.06669196

iii. **No Taxi ID and No Census Tract in the features**: in addition to the Taxi ID, *removing Census Tract does not have significant impact in the estimation of the Fare where the difference with the benchmark models is less than 0.2% for most of the models as well (Table 7).*

Table 7. Models retrained without the Taxi ID and Census Tract features

	Linear Regression		Random Forest		AdaBoost		GradientBoost		XGBoost	
	Metric	**Difference**	**Metric**	**Difference**	**Metric**	**Difference**	**Metric**	**Difference**	**Metric**	**Difference**
R2 Score	0.93751706	-0.02731641	0.99195750	-0.00065917	0.93456759	0.00213208	0.99141389	-0.00062080	0.99236992	-0.00074360
MSE	14.17934893	6.19895335	1.82509607	0.14958689	14.84867538	-0.48383663	1.94845990	0.14087980	1.73150574	0.16874640
RMSE	3.76554763	0.94058823	1.35096116	0.05654656	3.85339790	-0.06227727	1.39587245	0.05140970	1.31586691	0.06576318

iv. **No Taxi ID, No Census Tract, and No Community Area in the features**: further *removing the Community Area does not have significant impact in the estimation of the Fare where the difference with the benchmark models is less than 0.2% for most of the models too (Table 8).*

Table 8. Models retrained without the Taxi ID, Census Tract, and Community Area features

	Linear Regression		Random Forest		AdaBoost		GradientBoost		XGBoost	
	Metric	**Difference**	**Metric**	**Difference**	**Metric**	**Difference**	**Metric**	**Difference**	**Metric**	**Difference**
R2 Score	0.93171767	-0.03311580	0.99085689	-0.00175978	0.93335409	0.00091858	0.99007153	-0.00196316	0.99126065	-0.00185287
MSE	15.49541336	7.51501778	2.07485906	0.39934988	15.12405763	-0.20845439	2.25308279	0.44550269	1.98323450	0.42047516
RMSE	3.93642139	1.11146200	1.44043711	0.14602250	3.88896614	-0.02670903	1.50102725	0.15656449	1.40827359	0.15816986

v. **No Taxi ID, No Census Tract, No Community Area, No Tips, No Tolls, and No Extras in the features**: Finally, *removing Tips, Tolls, and Extras also shows that there is no significant impact in the estimation of the Fare with less than 0.5% accuracy loss with respect to the benchmark models (Table 9)*.

Table 9. Models retrained without the Taxi ID, Census Tract, Community Area, Tolls, Tips, and Extras features

	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

vi. **Conclusion regarding features**: in this analysis it was shown that most features in the dataset are not relevant in the estimation of the Fare, which only leaves ["Trip Seconds", "Trip Miles", "Payment Type", "Company"] as the only relevant fields that could impact the Fare significantly. In consequence, rather than using only samples inside of Chicago, I will combine all data points regardless whether they are inside/outside Chicago, in order to estimate the Fare of a ride.

3. Final Model

a. **Model selected**: from the different models tested Random Forest, GradientBoost, and XGBoost were the most successful, therefore I will proceed the final analysis with "**XGBoost**" as it is much faster than the other approaches thanks to its efficient implementation.

b. **Prediction Performance**: the only features to be considered for the training of the final "**XGBoost**" are ["Trip Seconds", "Trip Miles", "Payment Type", "Company"]. As for the amount of samples to be used, I have a total of 3,199,179 data points to use for training/testing. I will split the data into 80:20 ration for training and testing giving me the results shown in Table 10. It is clearly observed that ~97% of the variance from the dependent variable (Fare) can be described by just considering the 4 independent variables mentioned previously.

Table 10. Prediction results of "XGBoost" algorithm

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

c. **Feature Importance**: in terms for feature importance, the weights given by "XGBoost" model are shown in Figure 8, supporting the initial assumption that Trip Miles and Trip Seconds are indeed the most relevant features when estimating the Fare of a trip. To further show how strong the relationship between Fare and ["Trip Miles", "Trip Seconds"], the scatter plots are shown in Figure 9 for a limited amount of samples for visualization purposes. As for the ["Company", "Payment Type"], the weights are quite small suggesting that they could also be removed.

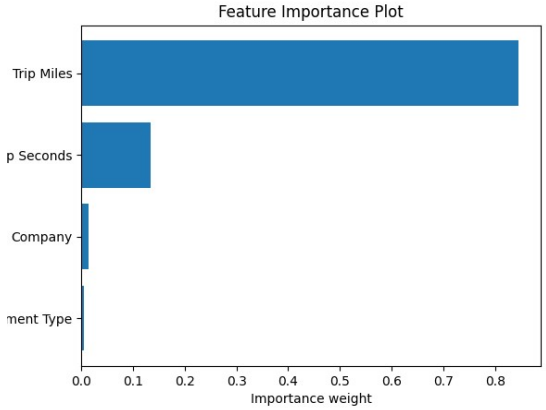

Figure 8. Feature importance coming from “XGBoost” model.

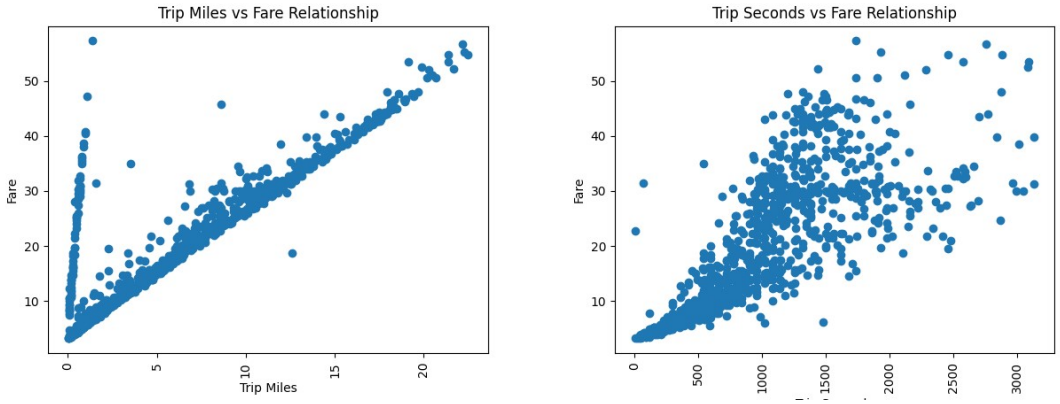

Figure 9. Scatter plots or Trip Miles and Trip Seconds against Fare (left, right respectively)