Uber Cool, Uber Pool

Kenta Yoshii, Colby Porter, Keigo Hachisuka, Grace Chen

Context

Ride-hailing services have gained popularity in recent years since it has reduced travel costs, traffic congestion, and emissions. With the rise of Uber, Lyft, and other ride-hailing services, the demand for ride-sharing services has also increased. Moreover, Covid-19 has brought a surge in ride-sharing demand. Our goal was to build models that could predict the price of ride-sharing services in Chicago based on the Covid-19 cases/deaths to see if there was any correlation between the two.

Data Set

We scraped two publicly available datasets provided by the Chicago Data Portal spanning November 2018 to February 2022

- Transportation Network Providers: captures all ride share trips in Chicago starting from November 2018
- COVID-19 Daily Cases, Deaths, and Hospitalizations: includes the number of deaths and daily cases per day

We joined these two datasets on timestamp. Our final dataset had Total Price as a target attribute and the eight other features including trip_miles and cases_total.

Methodology

- used principal component analysis to learn the distribution of our dataset
- created three different models to predict the prices of ride share data: logistic regression, KMeans + SVM, and LSTM
- used MSE to score the performance of different models between their predictions and the actual Uber prices in Feb 2022

Results

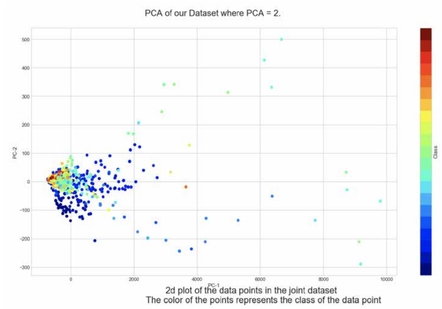

Figure 1. Used PCA to reduce the dimensionality of the graph, which showed us that the data was in one large cluster.

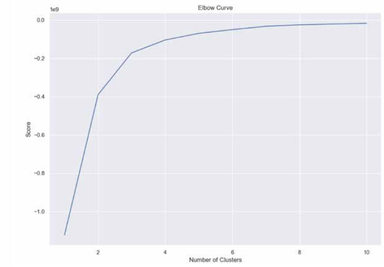

Figure 2. To figure out how many clusters to use for KMeans, we used an elbow curve, which showed us the optimal number was 3.

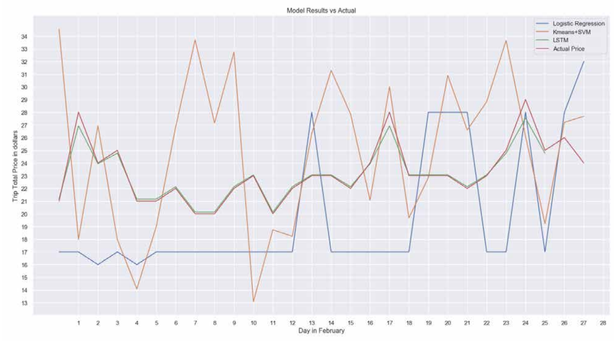

Figure 3. The predicted prices of each of the three prediction models along with the actual prices of the Uber rides.

Hypothesis Testing

Chi-Square Test for Independence

Test for whether Covid cases and trip price are independent of each other.

2 Test Statistic = 10705.839180
p-value = 0.0001
Critical Value = 10323.781777

Since p-value < 0.05, we reject the null hypothesis and conclude that Covid cases and trip price are not independent.

Augmented Dickey-Fuller Test

Test for whether trip_total is non-stationary/has a unit root.

ADF Statistic: -1.349072
p-value: 0.606434
Critical Values:
1% : -3.439, 5% : -2.865, 10% : -2.569

Since p-value > 0.05, we fail to reject the null hypothesis and cannot claim that our data is stationary.

Multivariate Normality Test

Test for whether our time-series data set has a Gaussian distribution.

p-value: 0.0012

Since p-value < 0.05, we reject the null hypothesis and conclude that our time series data set is not MVN.

Takeaways

Takeaway 1: The deep Learning model performed the best in predicting the prices for February. Table 1 shows that the LSTM model did an excellent job predicting the price as it handles time series data exceptionally well.

Model	MSE
Linear Regression (Baseline)	37.68
KMeans + SVM	25.42
LSTM	0.19

Takeaway 2: Each model captures different aspects of the dataset.

- Linear regression: sometimes captures the sudden spikes in the actual price
- KMeans + SVM: does a fairly good job capturing the general shape of the actual price, but at some points, it spikes more
- LSTM: almost identical to the actual price

Takeaway 3: Most of the predictive power comes from features such as "cases_total" and "trip_miles". The table on the right shows how our KMeans + SVM model performed with different subset of features.

Features	MSE
COVID cases	30.64
COVID Cases + hospitalization total + deaths total + cases age	29.14
All features	25.42

Conclusion + Next Steps

After visualizing the dataset, evaluating the clusters, and training three different models, we were able to achieve our goal – accurately predicting ride share prices based on the number of COVID cases and past ride-share data using an LSTM model. However, there is more work that could be done to make our work even more useful. For example, we could augment the model in the future to inflation rate and gas prices which are likely to have had a large impact. Additionally, future work could model cases and prices over months to see broader trends in the pandemic and pricing rather than a day-by-day analysis.