Data Glacier

Your Deep Learning Partner

G2M Case Study

Virtual Internship

Vincent Machoka Obiri

08-Jan-2022

Agenda

Executive Summary

Problem Statement

Approach

Data Exploration

EDA

EDA Summary/Recommendations

Machine Learning

Appendix

Data Glacier

Your Deep Learning Partner

Executive Summary

- Investing in any company necessitates a thoughtful, thorough and transparent business analysis to make an informed decision. XYZ, a private firm in the US, wants to invest in a Cab Industry that has seen exponential growth for the last few years. Particularly, the firm is interested in one of the two Cab companies- Yellow Cab or Pink Cab.
- Analysis such as Exploratory Data Analysis (EDA), feature engineering, statistics, and plots were performed on various cases on historical data from 2016 through 2018 for two companies.
- Based on results from these analysis, it is my recommendation to XYZ firm to invest in Yellow Cab company. Every analysis performed showed Yellow Cab company always performed better than Pink Cab.
- To further assist XYZ, a machine learning model was developed for Yellow Cab Company from 2016 data. The model prediction was validated using first quarter of 2017 data, which Model had never seen. The prediction accuracy range between 0.4% to 1.7%, which suggests a good model that can be deployed in production.

Problem Statement

- XYZ, a private firm in US needs to make a data driven decision to invest in a Cab company based on available datasets from two companies- Yellow Cab and Pink Cab.

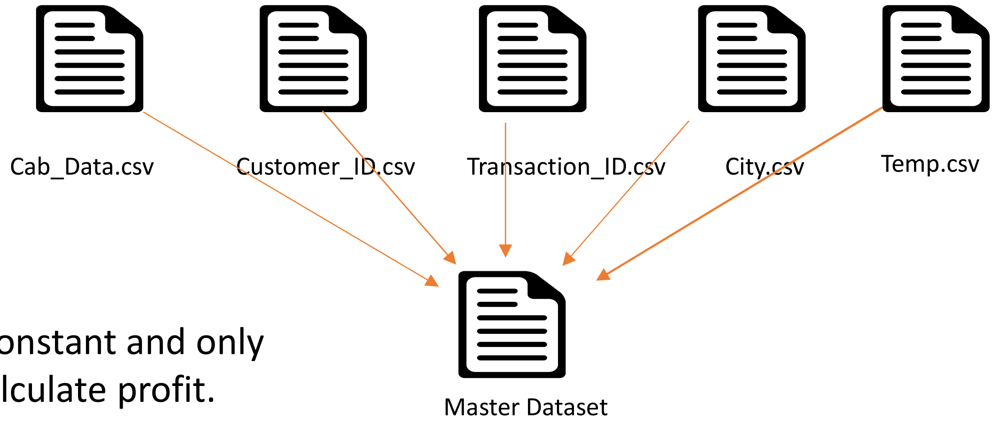
- The objective is to pick one company over another based on various data analysis.

Approach

- Historical Data collection from 2016 and 2018.
- Weather Data collection relative to Historical Data collection from climate.gov.
- Final Dataset created by combining Historical Data and Weather Data.
- Various Exploratory Data Analysis performed.
- Analyzed which Cab company performed better profit-wise.
- Built a machine learning model to help XYZ firm predict future profits.

Data Exploration

- Timeframe of the data: 2016-01-31 to 2018-12-31
- Total data points : 359,392



Assumptions:

- Profit of rides are calculated keeping other factors constant and only Price_Charged and Cost_of_Trip features used to calculate profit.

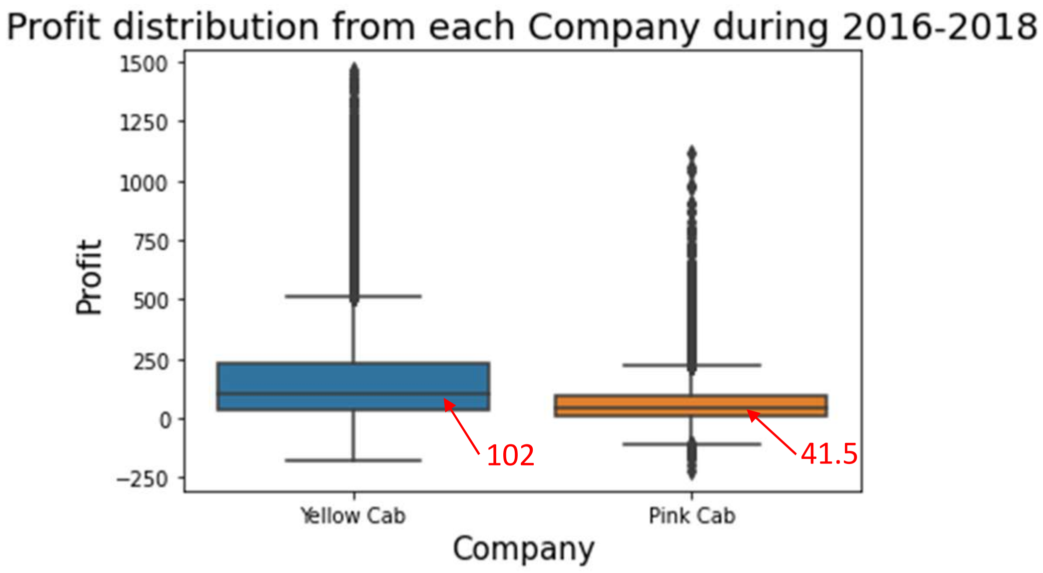
- Outliers are present in Price_Charged feature but due to unavailability of trip duration details ,we are not treating this as outlier.

- Users feature of city dataset is treated as number of cab users in the city. we have assumed that this can be other cab users as well(including Yellow and Pink cab)

Exploratory Data Analysis

- Five datasets Cab Data, Customer ID Data, Transaction ID Data, City Data and Temp_2016_2018 data were combined.
- Total of 16 features were present in final data set, and with 359,392 observations.
- One important feature target Profit was derived from Price Charged and Cost of Trip.
- All the Exploratory Data Analysis was performed with feature target Profit in mind.

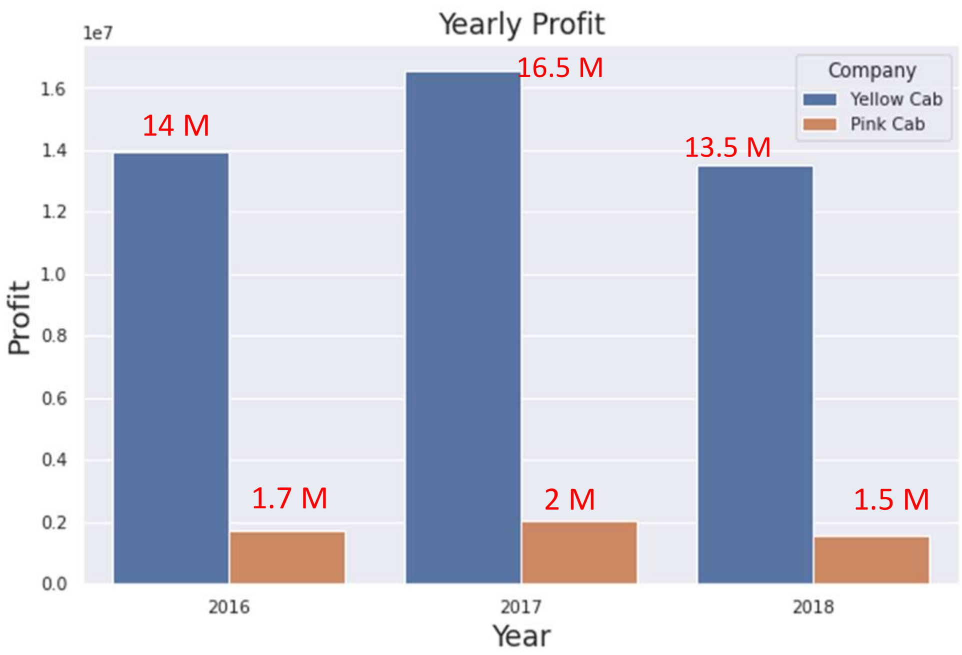
Box Plot Profit Distribution



- From the plot to the left, Yellow Cab did bring in more profit than Yellow.
- Median for Yellow Cab is 102.0 and for Pink Cab it is 41.5.

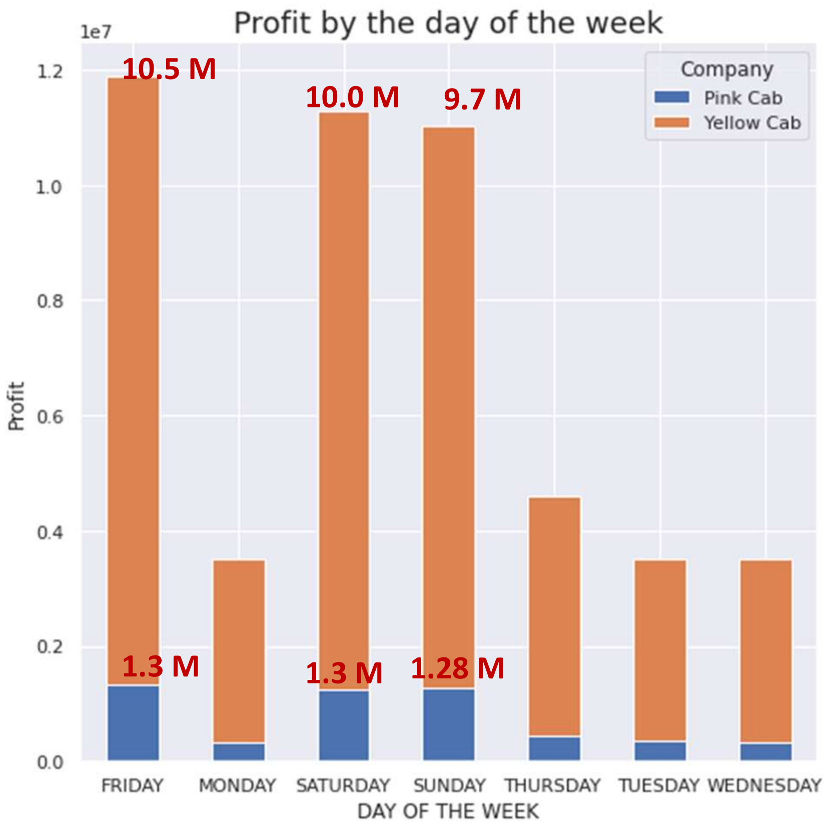
- Note: Machine learning model was fed all the above data including outliers that are seen here. The model however did not have any major issue, which tells these may be valid data!

Yearly Profit



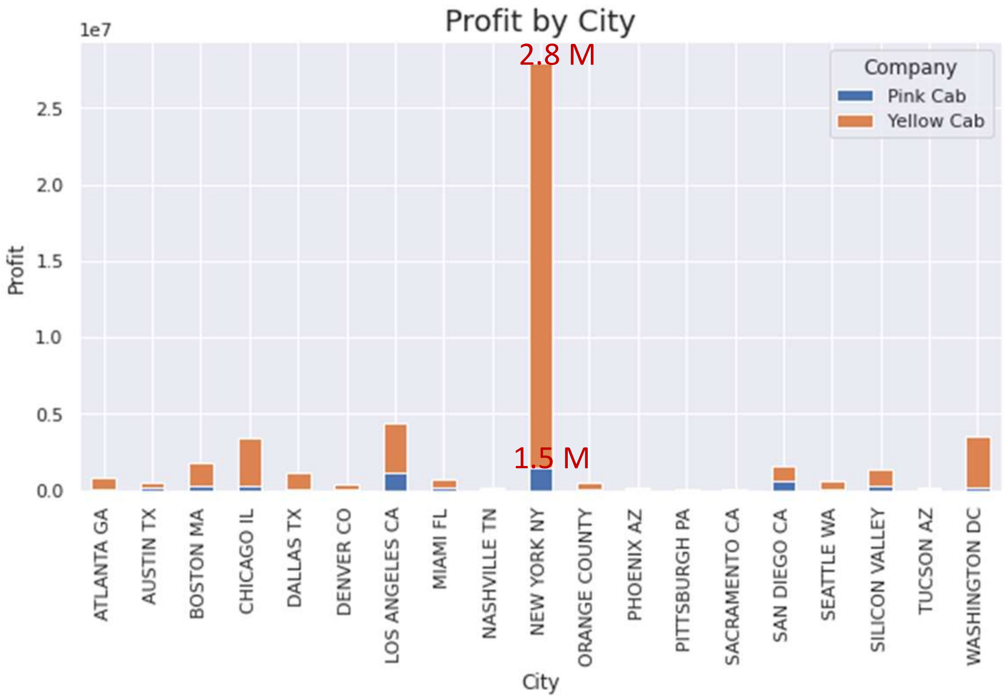
- Looking at Yearly Profit also, Yellow Cab brought in relatively higher profits than Pink Cab.
- In 2016, Yellow Cab brought 14 million whereas Pink Cab brough in only 1.7 million, and the same trend for 2017 and 2018 from the plot to the left.

Profit by the day of the week



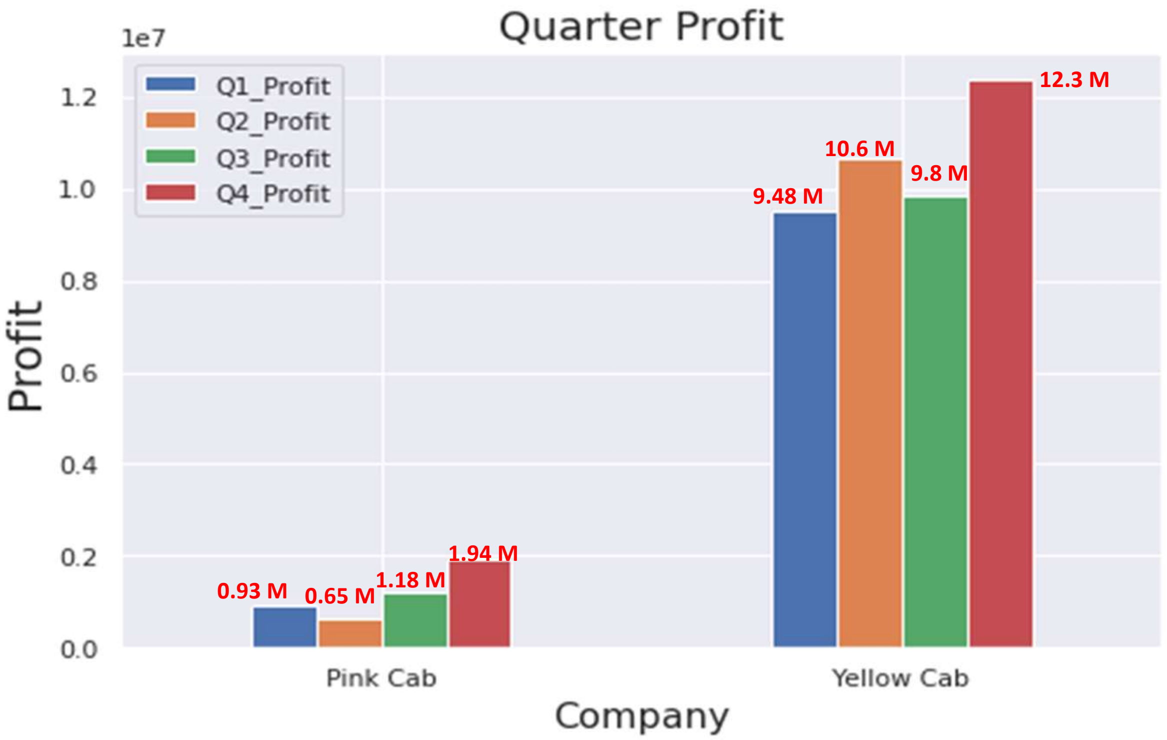
- Analyzing Profit by the day of the week during 2016-2018 also suggests Yellow Cab performed much better than Pink Cab.
- On Friday, Yellow Cab generated 10.5 million profit compared to 1.3M by Pink Cab, and similar trends follow for Saturday and Sunday.

Profit by City for each Company



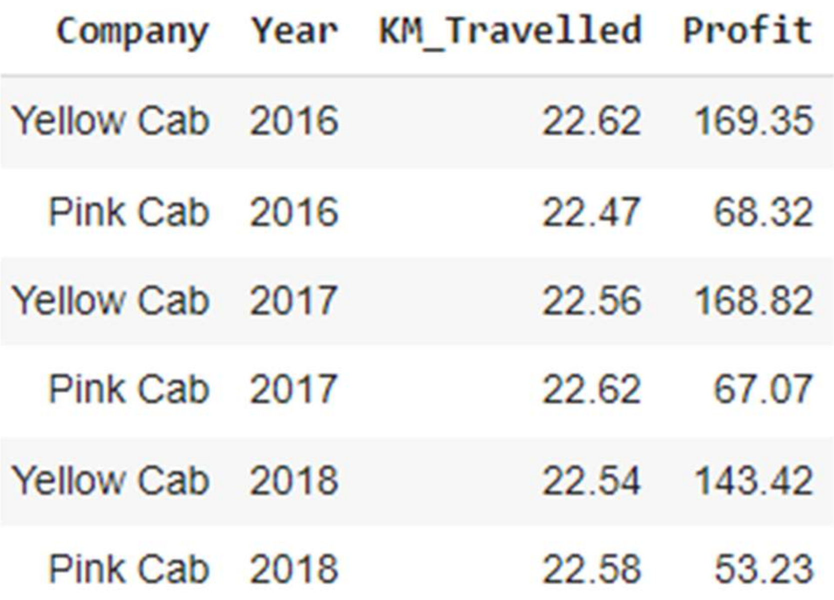
- In every city across, Yellow Cab performed better than Pink Cab.
- New York has more business for both companies. Yellow Cab brought in 2.8 M profit here compared to 1.5 M for Pink Cab approx.

Quarterly Profit 2016-2018



- Quarter data at each year (2016 thru 2018) calculated and combined.
- Yellow Cab company clearly outperforms Pink Cab in all quarters.
- Ex: Yellow Cab in Q4 brought in 12.3 million dollars compared to just 1.94 million by Pink Cab.

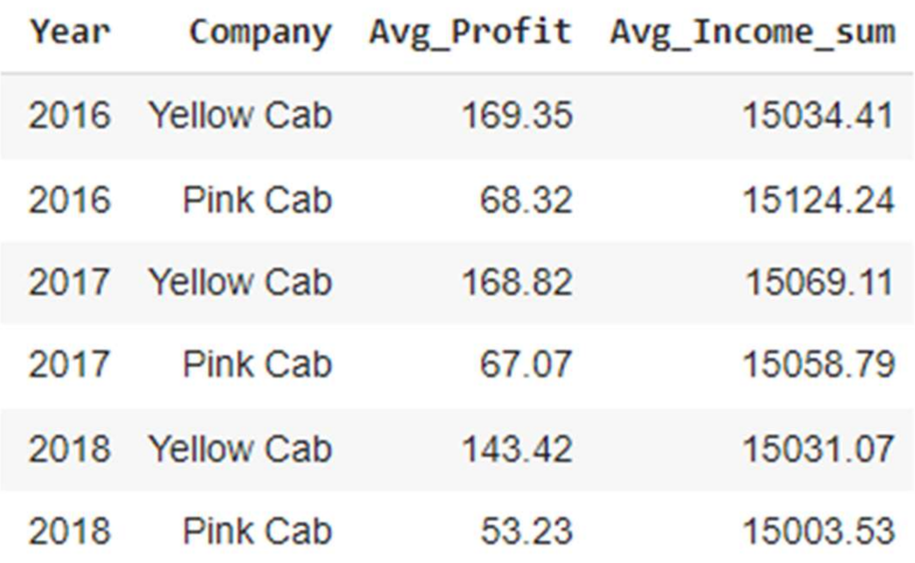
KM Travelled average vs Profit Average



Company	Year	KM_Travelled	Profit
Yellow Cab	2016	22.62	169.35
Pink Cab	2016	22.47	68.32
Yellow Cab	2017	22.56	168.82
Pink Cab	2017	22.62	67.07
Yellow Cab	2018	22.54	143.42
Pink Cab	2018	22.58	53.23

- Average Profit per average KM Travelled was also higher for Yellow Cab than Pink, as shown by a table to the left in each year. Yellow Cab is a clear choice again!

Average Income vs Profit Average



Year	Company	Avg_Profit	Avg_Income_sum
2016	Yellow Cab	169.35	15034.41
2016	Pink Cab	68.32	15124.24
2017	Yellow Cab	168.82	15069.11
2017	Pink Cab	67.07	15058.79
2018	Yellow Cab	143.42	15031.07
2018	Pink Cab	53.23	15003.53

- Average Profit per average income was also higher for Yellow Cab than Pink, as shown by a table to the left in each year. Yellow Cab is a clear choice again!

EDA Summary/Recommendations

- Data analysis such as overall Profit distribution, Yearly Profit, Profit at the day of the week, Profit by City, Quarterly Profit, KM Travelled average against Profit Average and Average Income against Profit Average all show Yellow Cab company is far more profitable than Pink Cab!
- Based on this assessment, I recommend XYZ firm to invest in Yellow Cab.
- To assist XYZ firm further, Machine learning model was developed for Yellow Cab data, which is presented in next slides.

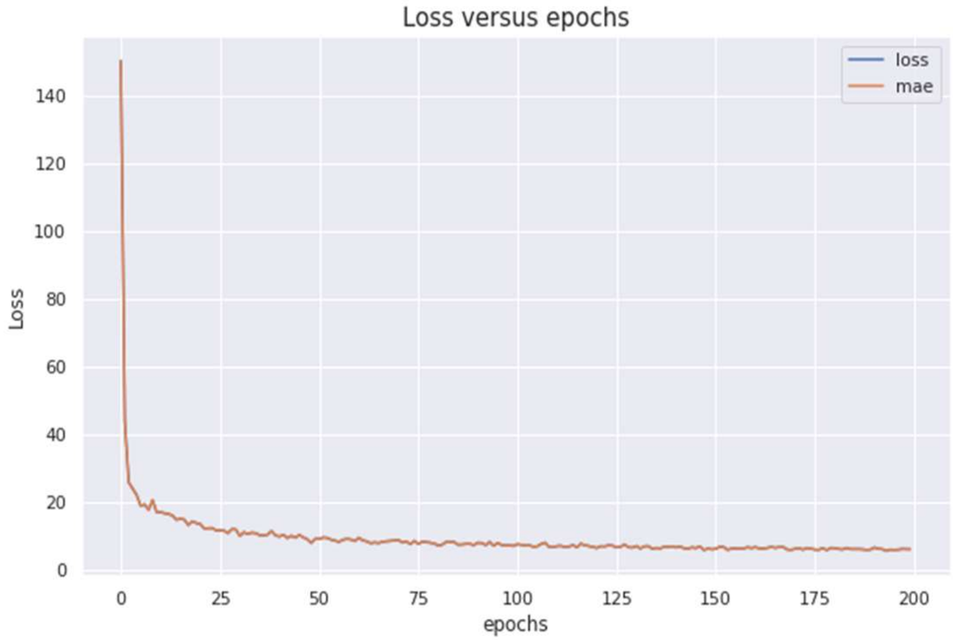
Machine Learning (ML) Model-Feature selection

- 2016 Yellow Cab data that are pre-processed prior to EDA.
- Feature selection performed based on Pearson correlation since we are predicting numerical variable from numerical features.
- Number of features considered based on Pearson correlation value closer to 0.5 or higher. Initially, features selected are KM Travelled, Price Charged, Cost of Trip, Population and Users. Correlation table is presented as an Appendix.
- Population and Users have a correlation factor of 0.92. Including Population has a feature in ML model actually made the model worse- hence dropped. More importantly profit is being generated by actual number of cab users.
- The final features are: KM Travelled, Price Charged, Cost of Trip and Users, to predict target variable Profit.

ML Model Data Distribution

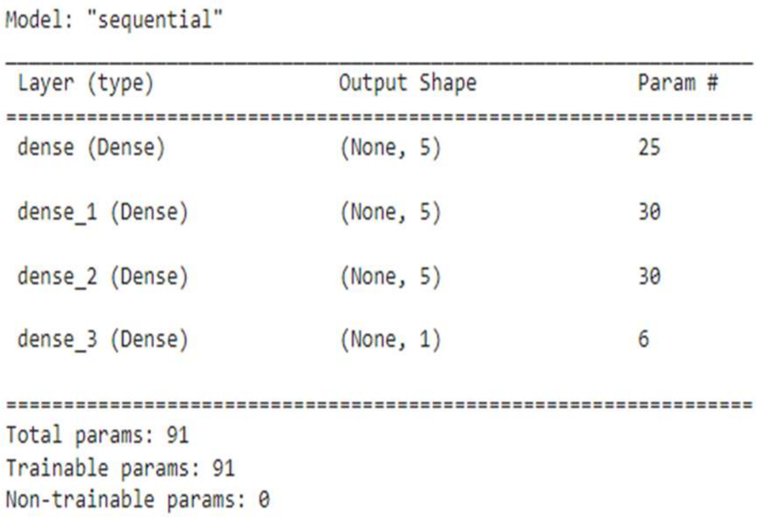
- Training set: 80% of dataset
- Test set: 20% of dataset
- ML Framework used: Scikit-Learn, TensorFlow.
  - Scikit-learn used to split the data
  - TensorFlow used to actually build the ML model
- Features: KM Travelled, Cost of Trip, Price Charged, Users
- Target Variable: Profit

ML Model Loss(MAE) over iterations



- Loss and MAE are the same for the model.
- MAE starts to flattens after 125 epochs(iterations) and achieve a constant MAE after 175 iterations.

ML Model Structure



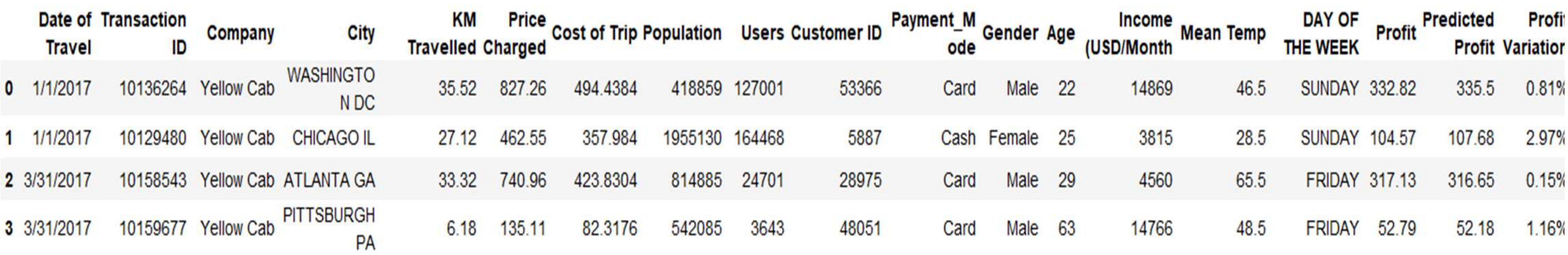
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 5)                 25

 dense_1 (Dense)             (None, 5)                 30

 dense_2 (Dense)             (None, 5)                 30

 dense_3 (Dense)             (None, 1)                 6

=================================================================
Total params: 91
Trainable params: 91
Non-trainable params: 0
```

- Sequential Deep Neuro Network (DNN) Linear Regression from TensorFlow utilized.
- Model Creation:
  - 3 Hidden Layers- each with 5 neurons with rectifier ReLU activation function.
  - 1 layer is the output with 1 neuron since we are predicting one Profit value each time.
- Model Compilation
  - Loss function of mean absolute error (MAE)- typical for linear regression
  - Optimizer: Adam optimizer with learning rate of 0.01031 for the best performance.
  - Metrics kept as MAE.
- Model Fit

  -80% training set, with epochs of 200.
- Model test

  -20% training set, resulted in MAE of 2.07

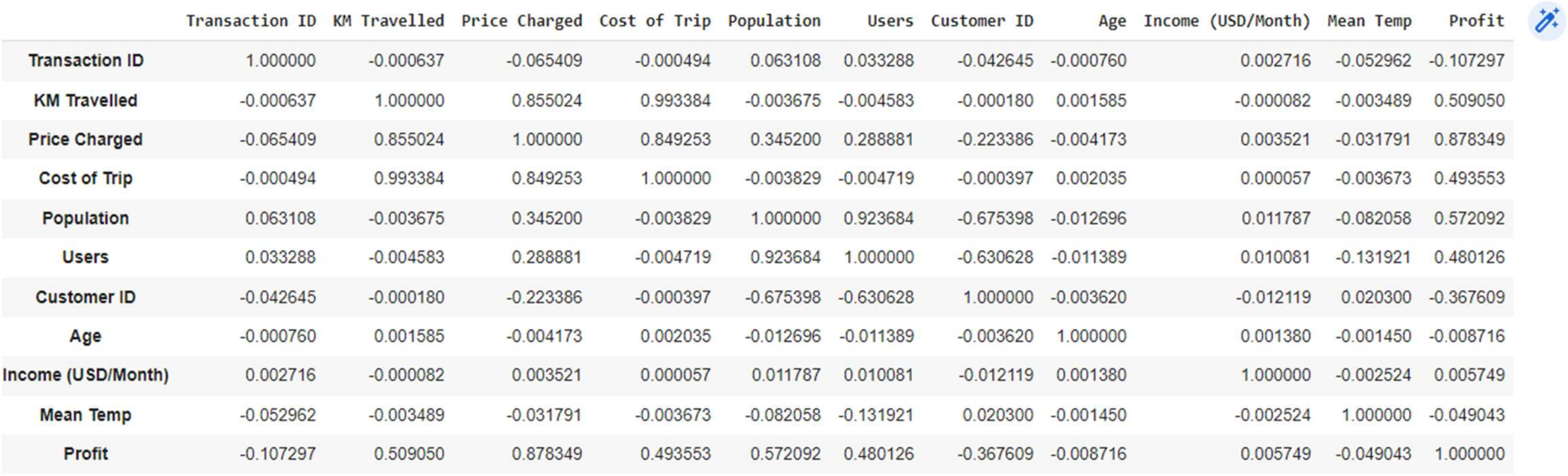
ML Model Validation



	Date of Travel	Transaction ID	Company	City	KM Travelled	Price Charged	Cost of Trip	Population	Users	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month	Mean Temp	DAY OF THE WEEK	Profit	Predicted Profit	Profi Variatior
0	1/1/2017	10136264	Yellow Cab	WASHINGTON DC	35.52	827.26	494.4384	418859	127001	53366	Card	Male	22	14869	46.5	SUNDAY	332.82	335.5	0.81%
1	1/1/2017	10129480	Yellow Cab	CHICAGO IL	27.12	462.55	357.984	1955130	164468	5887	Cash	Female	25	3815	28.5	SUNDAY	104.57	107.68	2.97%
2	3/31/2017	10158543	Yellow Cab	ATLANTA GA	33.32	740.96	423.8304	814885	24701	28975	Card	Male	29	4560	65.5	FRIDAY	317.13	316.65	0.15%
3	3/31/2017	10159677	Yellow Cab	PITTSBURGH PA	6.18	135.11	82.3176	542085	3643	48051	Card	Male	63	14766	48.5	FRIDAY	52.79	52.18	1.16%

- The table shows predicted profit from ML model next to real Profit data.
- The profit variation as shown in the table ranges from 0.15% to 2.97% for these randomly picked data from Q1 2017 data, which suggest ML model is really good.
- I recommend XYZ firm to start using this ML and optimize more if it chooses to.

Appendix



	Transaction ID	KM Travelled	Price Charged	Cost of Trip	Population	Users	Customer ID	Age	Income (USD/Month)	Mean Temp	Profit
Transaction ID	1.000000	-0.000637	-0.065409	-0.000494	0.063108	0.033288	-0.042645	-0.000760	0.002716	-0.052962	-0.107297
KM Travelled	-0.000637	1.000000	0.855024	0.993384	-0.003675	-0.004583	-0.000180	0.001585	-0.000082	-0.003489	0.509050
Price Charged	-0.065409	0.855024	1.000000	0.849253	0.345200	0.288881	-0.223386	-0.004173	0.003521	-0.031791	0.878349
Cost of Trip	-0.000494	0.993384	0.849253	1.000000	-0.003829	-0.004719	-0.000397	0.002035	0.000057	-0.003673	0.493553
Population	0.063108	-0.003675	0.345200	-0.003829	1.000000	0.923684	-0.675398	-0.012696	0.011787	-0.082058	0.572092
Users	0.033288	-0.004583	0.288881	-0.004719	0.923684	1.000000	-0.630628	-0.011389	0.010081	-0.131921	0.480126
Customer ID	-0.042645	-0.000180	-0.223386	-0.000397	-0.675398	-0.630628	1.000000	-0.003620	-0.012119	0.020300	-0.367609
Age	-0.000760	0.001585	-0.004173	0.002035	-0.012696	-0.011389	-0.003620	1.000000	0.001380	-0.001450	-0.008716
Income (USD/Month)	0.002716	-0.000082	0.003521	0.000057	0.011787	0.010081	-0.012119	0.001380	1.000000	-0.002524	0.005749
Mean Temp	-0.052962	-0.003489	-0.031791	-0.003673	-0.082058	-0.131921	0.020300	-0.001450	-0.002524	1.000000	-0.049043
Profit	-0.107297	0.509050	0.878349	0.493553	0.572092	0.480126	-0.367609	-0.008716	0.005749	-0.049043	1.000000

Data Glacier

Your Deep Learning Partner

Thank You