AN ANALYSIS OF NYC YELLOW TAXI'S:
PREDICTING OPTIMAL PICKUP LOCATIONS

Akira Takihara Wang
School of Computing and Information Systems
The University of Melbourne
Student ID: 913391
Tutorial: (Ruining Dong, Wednesday 6:15pm - 8:15pm)

September 28, 2019

ABSTRACT

The iconic Yellow Taxi's paint the life of New York - a city which never sleeps and is rapidly changing. With a total coverage of 50.8 million trips in 2017 and 2018 alone, the data provided by the Taxi & Limousine Commission (TLC) combined with ASOS Weather Observations will aim to address the effects of snowfall on a taxi drivers' profitability. The problem may be further extended to reducing traffic congestion, meeting passenger demands at the correct time, and may serve as an indication of vital roads during heavy snowfall.

Further exploration from the initial visualization phase yields a result which advises that highly populated zones will most likely have higher profitability, and is consistent regardless of year. Ultimately, if a taxi driver is in a local hot-spot (i.e a shopping district) at any given hour or day, then they will be profitable. If they are not, then a theorized model should be able to predict if they should stay in their current location or move to another location. The final model can predict with close to 75% classification accuracy when determining if a taxi driver is currently in an optimal location.

1 Building upon Phase 1 (Visualization)

The first phase [1] found external factors such as weather, time, roadworks and variable rates to affect the profitability of Yellow Taxi rides. Despite no conduction of statistical tests, there was clear evidence in the visualizations that suggested snowfall to be a factor impacting taxi rides. Other findings of notable interest included: decreased trip durations with no increase in frequency during weekday peak hours, a shortened trip duration but higher frequency during months with snowfall, and a higher estimated probability of tips when driving long distances. Since the data for snowfall is readily available, the analysis will work on the effects of snowfall on taxi rides.

For this reason, this report aims to predict optimal zone locations given the realistic environment of a taxi driver during their shift. These attributes include a taxi drivers' pickup location, time of pickup and potentially the current temperature and weather condition outside. The proposed theory is that there is much less demand in the outer suburbs during snowfall, and it is expected that the theorized model predictions indicate Central NYC / Manhattan and the Airports to be the most optimal zones during months with snowfall.

1.1 Analysis Assumptions

- The analysis works under the assumption that trip frequency in a zone $\approx$ taxi demand in a zone. This is due to the TLC Yellow Taxi data [2] recording trips without further information on the number of taxis in the zone.
- According to the TLC Taxi Fares [3], the fare calculation works at a variable rate:

  > "50 cents per 1/5 mile when travelling above 12mph **OR** 50 cents per 60 seconds in slow traffic or when the vehicle is stopped."

  This is unaccountable with the available data, hence, all profit rates in the analysis will be a crude approximation assuming linear distance and constant velocity.

1.2 Availability and Usage of TLC Yellow Taxi Data

The 2017 data will be used to approximate distributions to determine a good profitability metric, and serve as the train labels. Likewise, the 2018 data will be used as a development set as it is the proceeding year and should remain consistently similar with the 2017 set.

As such, the 2019 data will be used as a test set and aim to be predicted outputs of the model. If the hypothesis is true, then the model should be able to perform equally with both the 2018 development data and 2019 data. If so, then the theorized model predictions for a given taxi driver can be assumed to be correct during months with snowfall.

(Figure 1) shows the zone availability during months with snowfall in 2017, and by inspection, it is observable that the operable zones are much fewer, with Staten Island failing to make even a single trip per day. Complimenting is (Figure 2), which displays the number of snowfall warnings made during 2017. Both figures seem to show Central NYC / Manhattan and the airports to be both the busiest and most affected by snowfall. Logically, this makes sense since one would expect NYC Department of Transport to clear roads from snowfall that are used most, hence taxi trips may be limited to operating near central locations anyways.

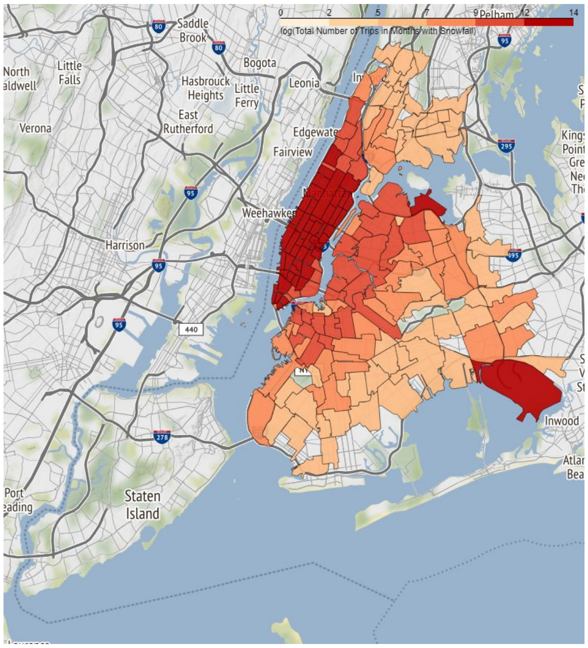


Figure 1: log(Total Number of Trips) in 2017 with at least one trip per day.

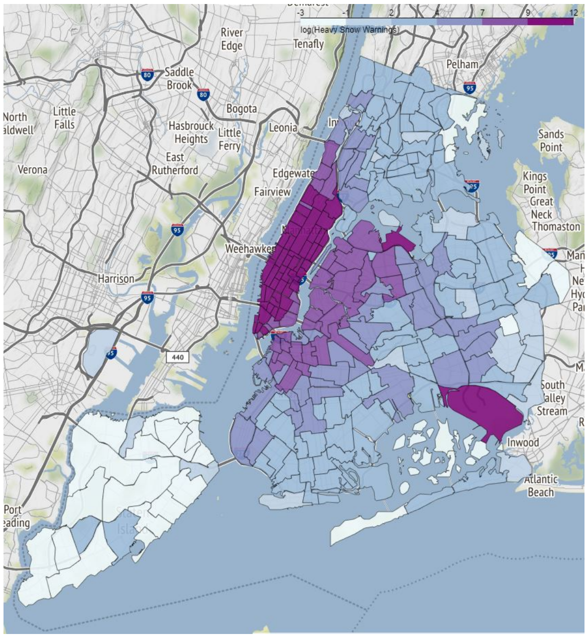


Figure 2: log(Total Snowfall Warnings) in 2017, where warnings are issued during intense snowfall conditions.

2 Preprocessing

The raw data mostly remains consistent with the provided Data Dictionary, and there is some indication that the data has been cleaned before released. Although the first phase [1] conducted a rigorous procedure of preprocessing, additional adjustments need to be made due to the nature of the second phase of analysis and predictions.

2.1 ASOS Weather Observations

Building upon the first phase [1], weather observations will now include a fourth station (LaGuardia Airport) after it was discovered that it hosted a large portion of pickups and dropoffs.

1. Missing values ("M") and Trace Amount ("T") were replaced with NaN's and imputed with zero under the recommendation of Daryl Herzmann - the data owner.
2. The weather code attribute "wxcodes" was categorized into the following attributes: "heavy_snow", "snow" or "normal_weather". This is because we are only interested in months with snowfall, and the attribute was deemed to be ordinal by inspection ("heavy_snow" is greater than "snow").
3. The "ice_accretion" attribute was dropped since it is not useful for the analysis as found in the first phase [1].
4. The "valid" attribute (validated timestamp) was renamed, converted and rounded to the closest standard DateTime hour. This is due to the precision of seconds and minutes is regarded as unnecessary in the context of this study, with further implications of increased space complexity if they were kept.
5. The newly created DateTime attribute was then categorized into the day of week and hour of the day.
6. A groupby concerning "datetime" was executed to unify the expected snowfall readings across the four weather stations. An exception to this was the temperature attribute, which had the maximum taken to denote the "worst-case" temperature of the day.

2.2 TLC Yellow Taxi Data

1. All values that are classified invalid according to the data dictionary are dropped instantly. These included trips that were negative in values even if they were correct since we are interested in predicting profitability, not the disputes.
2. Datetime attributes are categorized into the day of week and hour. Like the reasoning in the ASOS Weather Observations, the precision of minutes and seconds was unnecessary for the analysis
3. All "RatecodeID" with values 5 and 6 were dropped. These codes denoted Pre-Negotiated fares or Group Rides, which are the cause of several outliers.
4. Also, all trips without "payment_type" 1 or 2 were dropped. This is because the other values denote No Charge, Dispute, Unknown or Voided Trips - and are also the cause of several other outliers.
5. Finally, there is a minimum $2.50 hailing fee for a Yellow Taxi, so any instance with a "total_amount" less than this is incorrect and therefore dropped.

2.3 Data Merging

The two data sources were merged on a multi-index concerning the day of week and time of day. The resulting dataset is a cleaned TLC Yellow Taxi dataset where each trip has additional attributes indicating the maximum temperature, number of snow warnings and snowfall amount during the trip.

Since the dataset now reached well over 23GB as a CSV file, it was serialized into a feather format [4] and read in partitions during training and analysis.

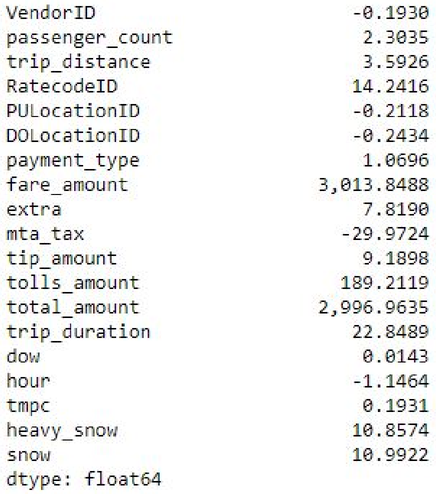
3 Attribute Analysis

3.1 Defining Posterior attributes from 2017 TLC Yellow Taxi Data

We define the posterior attributes of the dataset to be any information that relies on future information during taxi trips. These include the total fare, distance driven, time is taken, tips received and additional surcharges made. These attributes will be used to determine a zones' profitability, but will not be used when predicting a profitable zone during a taxi drivers' shift.

3.2 Skewness in Data Distribution

Skewness [5] is one way of measuring the asymmetry of an attributes distribution, where a high skew will usually result in a much higher mean over median - an unwanted property in this analysis when working with several instances.



```
VendorID                    -0.1930
passenger_count              2.3035
trip_distance                3.5926
RatecodeID                  14.2416
PULocationID                -0.2118
DOLocationID                -0.2434
payment_type                 1.0696
fare_amount              3,013.8488
extra                        7.8190
mta_tax                    -29.9724
tip_amount                   9.1898
tolls_amount               189.2119
total_amount             2,996.9635
trip_duration               22.8489
dow                          0.0143
hour                        -1.1464
tmpc                         0.1931
heavy_snow                  10.8574
snow                        10.9922
dtype: float64
```

Figure 3: Initial Attribute Distribution Skew with respect to a Normal Distribution.

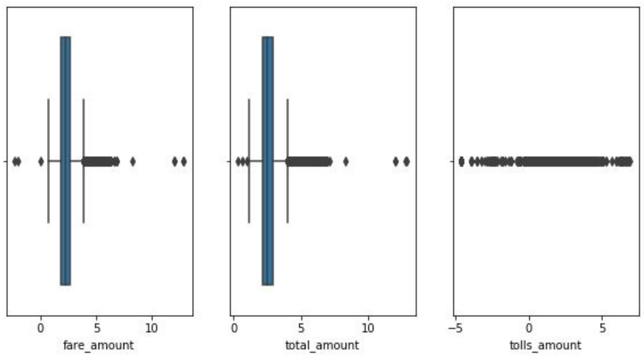


Figure 4: A log-scale box-plot of the Initial Attribute Distributions. It is visible that there are significant number of outliers that are affecting the distribution.

3.3 Initial Attribute Distributions

Even with preprocessing, (Figure 3) shows the skew of the current data and exposes significantly negatively skewed data. Of notable interest, "fare_amount", "tolls_amount" and "total_amount" have negative skews of (3013, 189, 2996) from an asymmetric Gaussian distribution - an expected outcome, as it was found in the first phase [1] (visualization) that there were simply higher counts of short duration trips when compared with longer duration trips. This indicates a large number of significant outliers in the data, and (Figure 4) is a complimenting box-plot which uncover these.

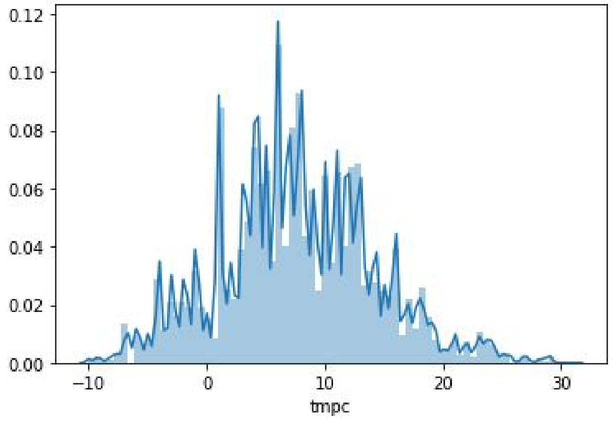


Figure 5: Distribution of Maximum Temperature.

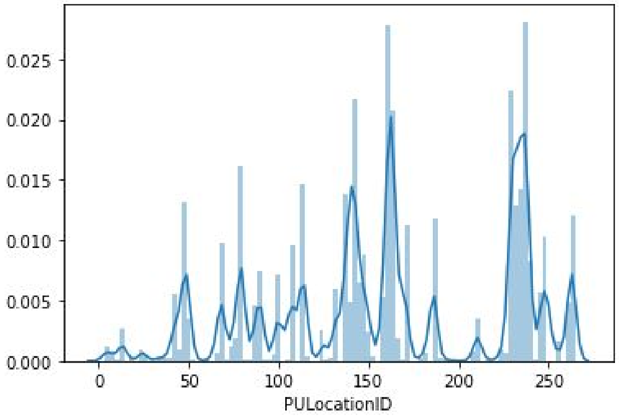
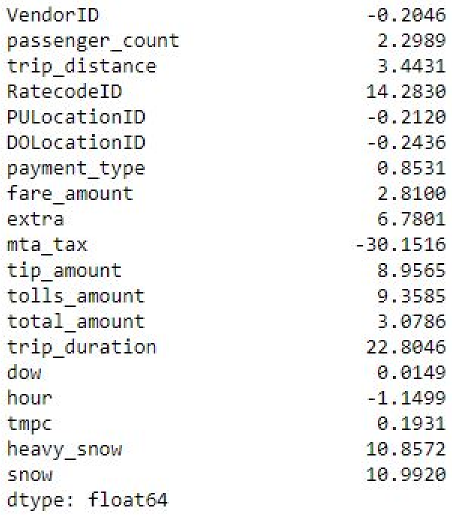


Figure 6: Distribution of Taxi Pickups

- From (Figures 3 and 4), the distribution is negatively skewed. A further look into the data suggests outliers with "fare_amount" as high as $4000 exists, which is very unrealistic and unlikely. As such, all instances with greater than 8 standard deviations away from the mean will be dropped.
- Since the data is conditioned on months with snowfall, the expected temperature should be in the single digits, with some days falling below sub-zero temperature. (Figure 5) shows the distribution of the data and seems to validate the claims.

- The distribution of the Pickup Locations from (Figure 6) shows several peaks, which is expected since Central NYC / Manhattan and the Airports are busier than the suburbs. Since the attribute is nominal, it does not need to be taken account and we can keep all instances.
- For the "tolls_amount" distribution (Figure 4) which has mean and all quantiles 0, any instance above $374.94 should be dropped. This is because even if a passenger passed through every tollway **and** paid for the return toll fee (prices are taken from [6]), the maximum cost would only reach $374.94. As a safe measure, any toll amount greater than this will be assumed incorrect and dropped.
- Also, a predictive model will require consistent data since it is looking to find the more optimal locations instead of predicting once-off profitable trips. As such, any instance with a fare cost over $500 was dropped.
- Overall, a total of 140,696 instances are removed from the 2017 train set and 130,001 from the 2018 development set. Combined, it is a measly 0.4% of the dataset removed.



```
VendorID              -0.2046
passenger_count        2.2989
trip_distance          3.4431
RatecodeID            14.2830
PULocationID          -0.2120
DOLocationID          -0.2436
payment_type           0.8531
fare_amount            2.8100
extra                  6.7801
mta_tax              -30.1516
tip_amount             8.9565
tolls_amount           9.3585
total_amount           3.0786
trip_duration         22.8046
dow                    0.0149
hour                  -1.1499
tmpc                   0.1931
heavy_snow            10.8572
snow                  10.9920
dtype: float64
```

Figure 7: Cleaned Attribute Distribution Skew with respect to a Normal Distribution.

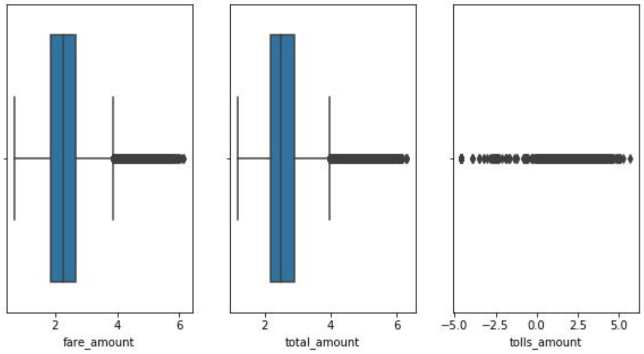


Figure 8: A log-scale box-plot of the Cleaned Attribute Distributions. Notice the x-scale is more tightly bound and there are much fewer outliers.

3.4 Cleaned Attribute Distributions

- (Figure 7) shows the new skew of the data, and it is instantly observable that it is much less skewed.
- "tolls_amount" is still negative skewed, but is now much more plausible under the realistic assumption that trips may go through several tollways.
- (Figure 8) also show more tightly bound instances, which will be useful when devising a profit metric.
- Attributes made redundant or not needed by the inspection were also removed:
  - "VendorID": The code which indicates the data provider and is not useful for calculating profitability.
  - "passenger_count": A redundant attribute since the "fare_amount" is unaffected by the number of passengers. For the group rides specifically, they were under "RatecodeID" 6 which were dropped.
  - "RatecodeID": The attribute has been conditioned on, so it can be dropped.
  - "payment_type": Like the "RatecodeID" attribute, it has been conditioned on and can be dropped.
  - "store_and_fwd_flag: The flag which indicates whether the data was held in the taxis' memory due to bad connection, which is not useful for calculating profitability.
  - "mta_tax": An attribute which is already taken into account by the "RatecodeID" according to [2].

3.5 Pearson Correlation Statistic

The Pearson Correlation [7] is one form of measuring the linear correlation between two attributes. The statistic is also not robust since it is affected by outliers and is quite sensitive to the distribution of attributes. Furthermore, a positive 1 correlation may signify an exact linear relationship (if variable X goes up, then variable Y will also go up), but must not be confused or misinterpreted as causality. This is merely one form of finding potential relationships and acts as an indication of possible relationships, not as a definitive cause and response.

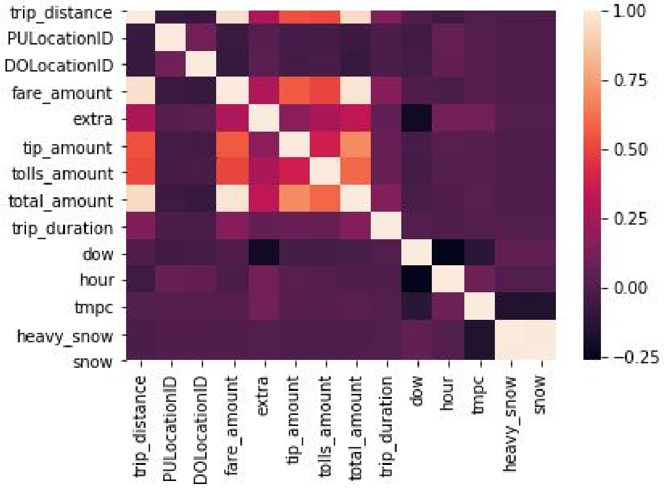


Figure 9: Correlation Heatmap of Relevant Attributes.

3.6 Discovering Inter-Attribute Relationships

- From (Figure 9), "trip_distance", "total_amount" and "fare_amount" all share a strong positive correlation as expected. In any case, "total_amount" will not be used since it incorporates several other factors which cannot be accounted for, and the analysis is focused on the *bare minimum* profit.
- "tip_amount" also shares a strong positive correlation with "fare_amount" (by default, the tip is 15% of the "fare_amount" [3]). If it is used as an attribute in profitability, the data must be conditioned on trips paying by card since cash tips were unrecorded. However, this will remove too many instances thus it will not be used.
- "tolls_amount" also seems to have a near strong correlation with both "fare_amount" and "trip_distance". Perhaps this attribute will be useful in devising some form of profit metric.
- Evidently, "snow" and "heavy_snow" share a near 1 correlation since "snow" is a subset of "heavy_snow". These attributes seem redundant but will be used for the predictions since they are prior attributes (a variable a taxi driver will know before trips).
- "trip_duration" seemed to not have much of a correlation with any other attribute, but in reality, one will know a longer trip will also have more "fare_amount" if the variable rate fare is ignored (as mentioned in assumptions).
- Although it may seem as if the "hour" attribute should be one hot encoded, the first phase [1] suggested that there were some dependencies of demand to the hour of the day. Hence, it is expected that during feature selection that there will be some form of a split between morning and afternoon.
- Since the feature space is quite sparse, there is no need for PCA or other high-level dimension reduction algorithms to be applied on this set of attributes.
- Finally, the most useful posterior attributes are as follows: "trip_duration", "trip_distance", "fare_amount" and "tolls_amount". In addition to this, an estimate for demand using frequency should also be relevant.

4 Generating Labels for Predictions

4.1 What makes a trip profitable

As observed in phase one [1], months with snowfall saw increases in trip frequency albeit shorter trip durations when compared to months without snowfall. Trips that were deemed profitable originated from pickup zones with high demands, as well as additional surcharges at play for the majority of trips.

The zones' profitability were calculated as:

1. Using time as a metric

$$\text{Zone Profit}_t = E\left[\frac{\text{Fare Amount} + \text{Tip Amount}}{\text{Trip Duration (minutes)}} \times \frac{\text{Frequency of Trips in Zone}}{\text{Total Number of Trips}}\right] \tag{1}$$

2. Using distance as a metric

$$\text{Zone Profit}_d = E\left[\frac{\text{Fare Amount} + \text{Tip Amount}}{\text{Trip Distance (Miles)}} \times \frac{\text{Frequency of Trips in Zone}}{\text{Total Number of Trips}}\right] \tag{2}$$

This analysis will take into account the suggested future considerations from phase one by accounting for the hour of day and day of the week to the zone. Furthermore, the median for the Total Number of Trips concerning zone instead of the mean will be utilized since it is less affected by outliers.

Hence, a new rating that defines taxi drivers potential profits using a crude approximation of demand can be devised as:

$$\text{Trip Rating} = \left[\frac{1}{2}\left(\frac{\text{Fare Amount}}{\text{Trip Duration}} + \frac{\text{Fare Amount}}{\text{Trip Distance}}\right) + \text{Mode Surcharge} + \text{Mode Toll Fee}\right] \times \frac{\text{Number of trips in Zone}}{\text{Median(Total Number of Trips)}}, \tag{3}$$

where the number of trips is with respect to the day of week, hour of day and pickup location. The reason why the average between "fare_amount" over "trip_duration" and "trip_distance" is taken, is due to the inconsistencies between the variable rates (short duration but super long distance or long duration over short distances). By taking the average, the values are smoothed over to give a better estimate of the true value. (Figure 11) justifies this claim by showing the fit of the final distribution.

4.2 The Proposed New Rating System

The new rating system is scaled by the proportion of trips inside the zone. This means that if a single substantially profitable trip only occurred once every few weeks, then the pickup zone is not as profitable. The assumption is that a taxi driver **will prefer** a more steady income. Subsequently, the most common toll fees and surcharges are also taken due to the expectation that the trip a taxi driver will take will be the "average" or "most common" trip. A prime example of such a trip is a morning trip from the suburbs into the city, which passes through a few tollways and pays for an additional morning peak hour fee.

(Figure 10) exhibits the distribution of the new rating system, which is certainly undesirable. On the other hand, to model a profitable trip only requires the relationship - that is, the current location and trip will be considered profitable if it is profitable when compared to other trips in the current hour. Therefore, a logarithm transformation will be applied to the distribution, resulting in the distribution depicted by (Figure 11).

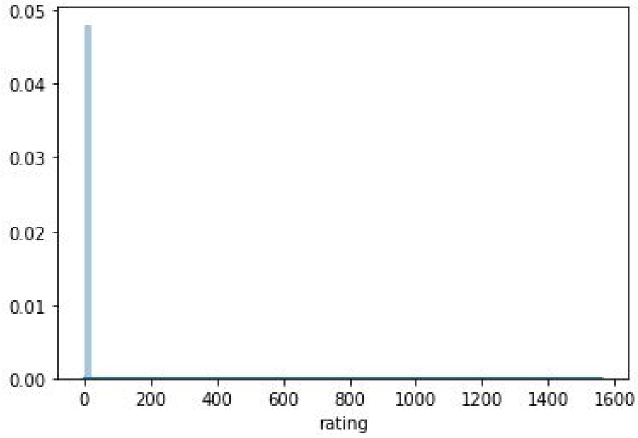


Figure 10: A Non-Scaled Distribution of the new Rating System. This is quite unwanted and undesirable for many reasons.

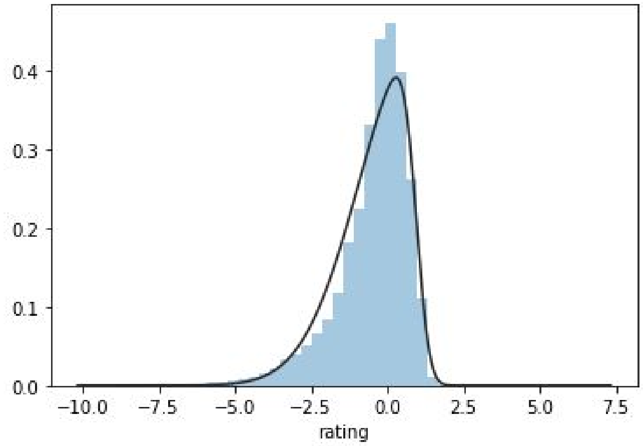


Figure 11: Logarithmic Transformed Distribution of the New Rating System, with a fitted line plot of the Skew Normal Distribution ($\alpha = -4$).

The log-transformed distribution looks significantly better even if it's negatively skewed. Astonishingly, it identifies approximately as a Skew Normal Distribution with shape parameter $\alpha = -4$. This is useful for accurately predicting profitable trips since it follows a known probability distribution, allowing a Generalized Linear Model (GLM) with a Skew Normal Distribution link function to be fitted.

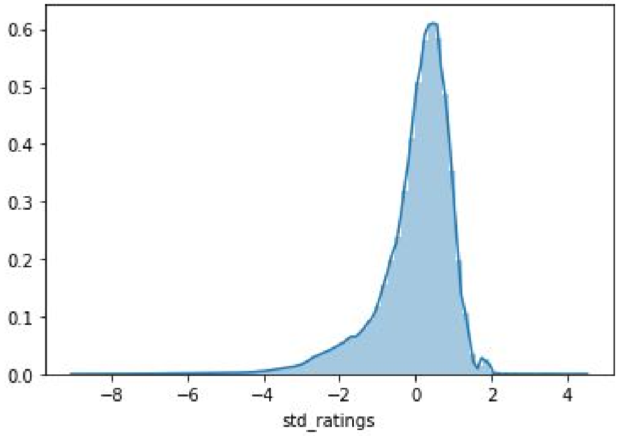
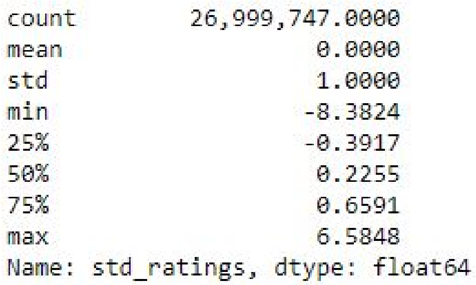


Figure 12: Plot of the final distribution of Rating after standardizing.

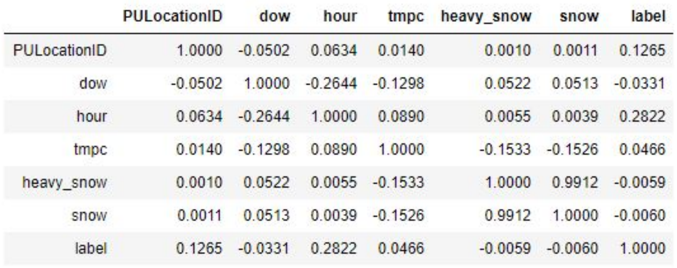


```
count       26,999,747.0000
mean                 0.0000
std                  1.0000
min                 -8.3824
25%                 -0.3917
50%                  0.2255
75%                  0.6591
max                  6.5848
Name: std_ratings, dtype: float64
```

Figure 13: Five Number Summary of the final distribution of Rating after standardizing.

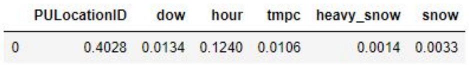
However, the time complexity of fitting such a model combined with limited computational resources means that it is unfeasible for this analysis. Thus, the attribute will be scaled to have mean 0 and standard deviation 1 to make it closer to a Gaussian as a bare minimum. (Figure 12 and 13) portray the final standardized distribution of rating.

4.3 Labelling the Trips



	PULocationID	dow	hour	tmpc	heavy_snow	snow	label
PULocationID	1.0000	-0.0502	0.0634	0.0140	0.0010	0.0011	0.1265
dow	-0.0502	1.0000	-0.2644	-0.1298	0.0522	0.0513	-0.0331
hour	0.0634	-0.2644	1.0000	0.0890	0.0055	0.0039	0.2822
tmpc	0.0140	-0.1298	0.0890	1.0000	-0.1533	-0.1526	0.0466
heavy_snow	0.0010	0.0522	0.0055	-0.1533	1.0000	0.9912	-0.0059
snow	0.0011	0.0513	0.0039	-0.1526	0.9912	1.0000	-0.0060
label	0.1265	-0.0331	0.2822	0.0466	-0.0059	-0.0060	1.0000

Figure 14: Correlation Table of the Realistic Attributes.



	PULocationID	dow	hour	tmpc	heavy_snow	snow
0	0.4028	0.0134	0.1240	0.0106	0.0014	0.0033

Figure 15: Mutual Information of the Realistic Attributes with respect to the Label.

The Rating will be binned into 5 equal frequency bins (two standard deviations away from the mean and median on both sides) to make this a multi-class prediction rather than a continuous prediction. The reasoning of doing is due to the assumptions of crude demand estimates, so, any continuous prediction (such as money per minute or distance) will already be an incorrect estimate of profitability. As such, the predictions of how profitable a trip can according to the new rating system be put into bands:

0. Rating (-inf, -0.611]: Not profitable at all (low frequency, low fare)
1. Rating (-0.611, 0.0248]: Not profitable (low frequency, medium fare)
2. Rating (0.0248, 0.396]: Neutral (medium frequency, medium fare)
3. Rating (0.396, 0.759]: Quite Profitable (high frequency, low fare / medium fare)
4. Rating (0.759, inf): Very Profitable (high frequency, medium fare / high fare)

It could also be said that a Taxi Driver will only have prior knowledge of a trip (such as the current location, weather and time), so only the realistic prior attributes will be retained when predicting. (Figure 14) is a correlation table of all the attributes, and (Figure 15) is the Mutual Information of the attributes concerning the label. There seem to be no signs of highly correlated attributes with the label, and the Mutual Information suggests that the "PULocationID" or "hour" may be a useful attribute when determining a split for tree-based models. These properties of an attribute to label relationship are sought for, so we can be content with these labels. They will also be applied to the months with snowfall in 2018 since we initially hypothesized that the profitable zones should be consistent regardless of year.

5 Predictive Models

The multi-class predictive modelling problem prevented the usage of common binary classifiers and continuous predictors such as Support Vector Machine (SVM) and Linear Regression (Logistic, Poisson, Binomial). Hence, considered models were required to have a multi-class counterpart, such as the Broyden–Fletcher–Goldfarb–Shanno (bfgs) algorithm for Logistic Regression.

Besides, the time and space complexities of several state-of-the-art models were unfeasible with the given computational power, so algorithms which could compromise between accuracy and time complexity were more favoured.

As such, a One-Rule / Decision Stump was used to give a baseline of the task at hand. Furthermore, due to the nature of the dataset with a mix of continuous, ordinal and nominal attributes, the extended Decision Tree and Random Forest models will be experimented on top of the multi-class Logistic Regression.

5.1 Evaluation and Models

5.2 Classification Accuracy, Precision and Recall

- **Classification Accuracy** is equal to the proportion of instances for which we have correctly predicted the label, and is defined as:

$$\text{Classification Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}. \tag{4}$$

- **Precision** is defined as the accuracy of predicting each class label, and is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{5}$$

- **Recall** is the proportion of correctly predicted class labels, and is defined as:

$$\text{Precision} = \frac{TP}{TP + FN}. \tag{6}$$

Due to the multi-class distribution, the macro averaged precision and recall will be taken.

5.3 Gini Impurity

Gini Impurity is a subset of Information Theory mainly used with Tree-based models as an evaluation metric. It is formally defined as the probability of an incorrectly classified instance:

$$G = \sum_{i=1}^{C} \mathrm{P}(i)(1 - \mathrm{P}(i)), i \in C, \tag{7}$$

where C is the class distribution.

A Gini Impurity of 0 will indicate a perfect split of the data, so the evaluation will be used for the Decision Tree (DT) and Random Forest (RF).

5.4 One-Rule (1R)

The One-Rule / Decision Stump model uses an algorithm which generates a single rule for each predictor in the data, and select the rule with the smallest total error as its "one rule".

This simple baseline will be used to give a *evaluative estimate* of the difficulty of the problem at hand within a reasonable time frame.

5.5 Multi-Class Logistic Regression (LR)

Logistic Regression is a Generalized Linear Model (GLM) which uses a logit function to estimate the probability of an instance and associate it with its binary labels.

It is regarded as a popular choice amongst researchers and is a contrast to the Tree-based models. Since our problem is Multi-Class, a multinomial variant using a limited memory bfgs solver will be utilised.

The evaluation of this model will be determined by the Precision, Recall and Classification Accuracy.

5.6 Decision Tree (DT)

An extended variant of the One-Rule, the Decision Tree will have as many rules for as many predictors that are allowed and will predict values by traversing down to a leaf node. The splits will be determined with either maximising Information Gain or Gini Gain, an information metric of determining the "goodness" of a split.

Since the data is quite skewed and we wish to retain some level of model interpretability, a minimum of 5000 samples per leaf node and a maximum depth of 5 will be enforced. The DT should always outperform the One-Rule due to its increased depth and decision capabilities.

In addition to the same evaluation for LR, the DT will also be evaluated by the Gini Impurity.

5.7 Random Forest (RF)

The Random Forest is a stacked ensemble learner which will predict based on the combination of several randomly generated Decision Trees as its base models.

By definition, the Random Forest will reduce the model variance and bias due to the numerous base classifiers, hence it is expected that the RF will outperform all of the models above.

The RF will use the same evaluation method as the DT.

5.8 Training and Development Stage

As mentioned initially, the 2017 data will be utilized as the train set for predictive modelling whilst the 2018 set will be used to develop and test each model's performance. The final goal is to give predictions for 2019 and recommendations for profitable locations. Furthermore, the DT and RF will have a limited depth of five to be interpretable by inspection and to see the general feature importance during training.

All training and predictions are done using the same Cloud Computing Instance hosted by Azure.

Model	Accuracy	Precision	Recall	$\approx$ Runtime
1-R	29.25%	15%	29%	2 min 48 sec
LR	29.25%	26%	29%	17 min 24 sec
DT	44.99%	49%	45%	4 min 23 sec
RF	41.99%	43%	42%	14 min 17 sec

Table 1: Initial Model Evaluation.

(Table 1) presents a table of evaluated results, with a poor accuracy of close to 30% for the 1-R model. Surprisingly, the LR performs just as poorly as the baseline model albeit a slight increase of Precision. However, with the *longest* train time of 17 minutes, the LR model fails to be competitive enough in both accuracy and time complexity when compared with the DT and RF. Although LR performed poorly, it can be attributed to the bad choice of attributes (i.e linearly inseparable data) - which the DT and RF were able to capture. This is due to the time attributes being treated as continuous, rather than categorical (i.e one hot encoded into separate attributes). To support the usage of continuous-time attributes over one hot encoded methods, tests were run using one-hot encoding and saw unjustifiable increases in time and space complexity (an additional 31 attributes to be created) without any significant increase in evaluation metrics. Therefore, it was decided that the time attribute should be continuous, and the LR will not be used in the final predictions.

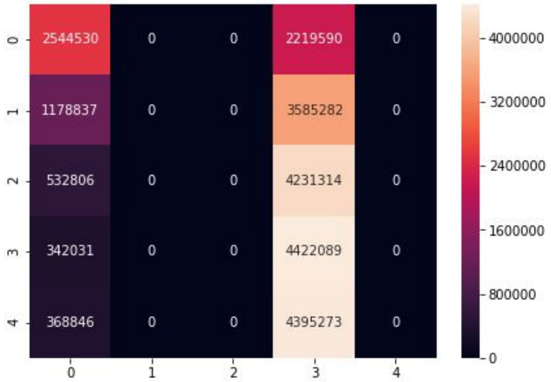


Figure 16: One Rule Confusion Matrix.

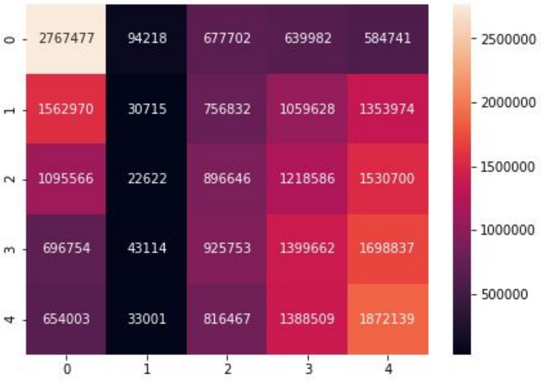


Figure 17: Logistic Regression Confusion Matrix.

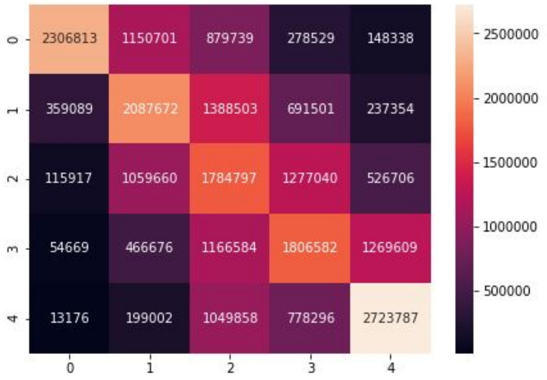


Figure 18: Decision Tree Confusion Matrix.

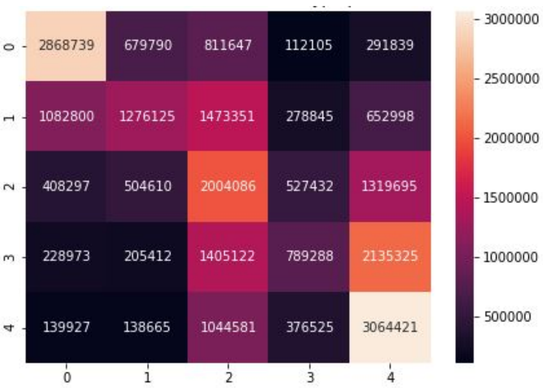


Figure 19: Random Forest Confusion Matrix.

6 Initial Stage - Error Analysis

By inspection of (Figures 16 - 19), it seems the class labels 1 - 3 are more difficult to predict in comparison to labels 0 and 4. Granting the aim of the analysis is to predict the profitability of locations, the outcome is acceptable since it will be at a bare minimum - able to discern differences between zones that are defined as "Not profitable at all" and zones that are "Very profitable". Out of the three non-baseline models, LR seems to suffer the highest inconsistency when predicting whilst the DT seems to be more consistent in terms of predictions - that is, the DT excels at predicting instances within one level of their true class. Although the RF shows to be much more accurate with labels 0, 2 and 4, the time complexity scales by an additional $\mathcal{O}(nm)$, where n is the number of base estimators and m is the number of samples per node. Therefore, the DT seems to be the best compromise in terms of performance and time complexity.

The DT and RF are the most consistent with class predictions but suffer from high variances with label 2 predictions. Label 2 is defined as the "Neutral" class - a class consisting of a mix of mediocre frequencies and varying fare rates. There is also a possibility of overfitting in the DT, although the learning rate from (Figure 20) suggests that it is not overfitting as more training instances are used. Even when compared with the LR learning curve (Figure 21), the DT convergence is quite stable and has a much lower error rate (calculated as 1 - Score) and much more accurate.

Average feature importance is also shown in (Figure 22), supporting the claims made by (Figure 14 and 15). The weather attributes have no feature importance during training as intended since the data is *already conditioned* on the months with snow, so the weather attributes shouldn't be describing any additional information. If any weather attribute *had* described additional information coincidentally, then the whole attribute selection would require revision. The Gini Impurity for both models exceeds 0.5 for the majority of leaf nodes - a behaviour akin to a random classifier. This implies that the DT has leaf nodes that are a mixture of classes and will resort to a random guess between classes at leaf nodes. However, this is still expected behaviour due to the max depth constraint of five.

The conclusion made the initial error analysis is that the DT is the best model when considering the compromise of time complexity and accuracy of predictions including precision and recall. As such, the final predictions for 2019 will be made using a Decision Tree.

6.1 Interpretation of the Decision Tree

The initial split from (Figure 22) uses the hour of day attribute - separating the instances to be either before 3 pm or after 4 pm (recall the trips are rounded to the closest hours). This is quite self-explanatory as the first phase [1] already suggested that trips made after 4 pm followed a different demand and distribution. Following the first split, the child branches split for pickup locations - successfully capturing the relationship of nearby locations together (i.e zones approximately surrounding Central NYC / Manhattan are split off together).

Since the Gini Impurity is dominantly large in each leaf node, a potential improvement is to prune any branch or node with an impurity greater than 0.5 to generate purer leaves. Additionally, the weather attributes can now be dropped after confirmation from the Feature Importance (Figure 21). More potential improvements may be seen by tuning hyper-parameters using Randomised Search (instead of Grid Search which is unfeasible for this large data set), however, it must be executed with caution to ensure no overfitting when doing so.

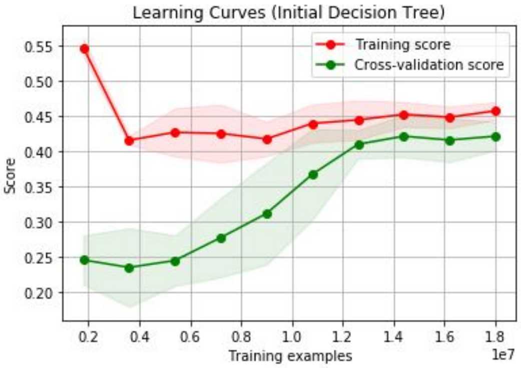


Figure 20: Decision Tree Learning Curve.

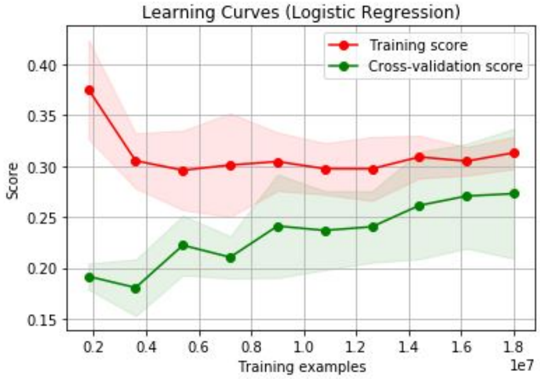


Figure 21: Logistic Regression Learning Curve..

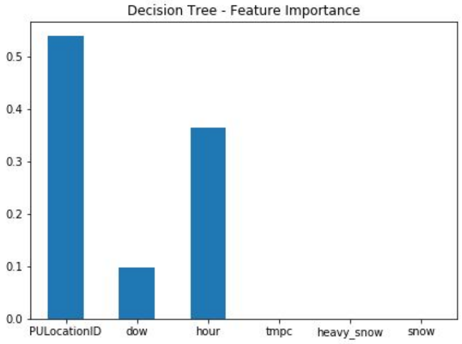


Figure 22: Average Feature Importance from the DT and RF.

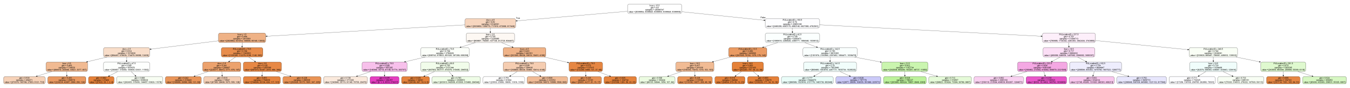


Figure 23: Overall Interpretation of the Initial Decision Tree with a minimum of 5000 samples per leaf node and a maximum depth of 5. (Darker coloured nodes have a higher Gini Impurity).

7 Final Results and Predictions

The optimal hyperparameter values using Randomized Search were discovered to be:

- A minimum sample of 250 instances per split and 5000 samples per leaf
- Unlimited number of leaf nodes (this ensures the purest of leaves)
- Maximum depth of 35 to prevent overfitting
- Using Gini instead of entropy as the split criterion.

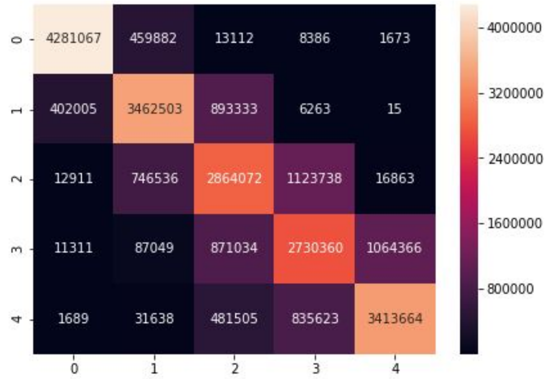


Figure 24: Final Decision Tree Confusion Matrix.

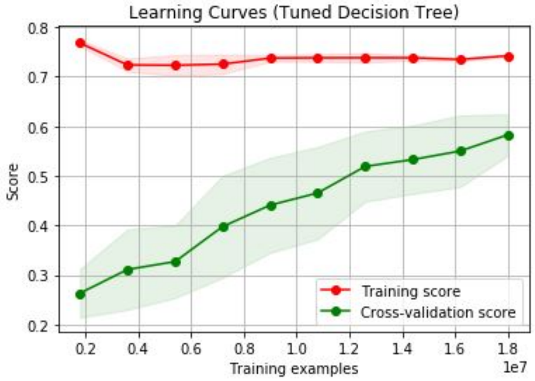


Figure 25: Final Decision Tree Learning Curve.

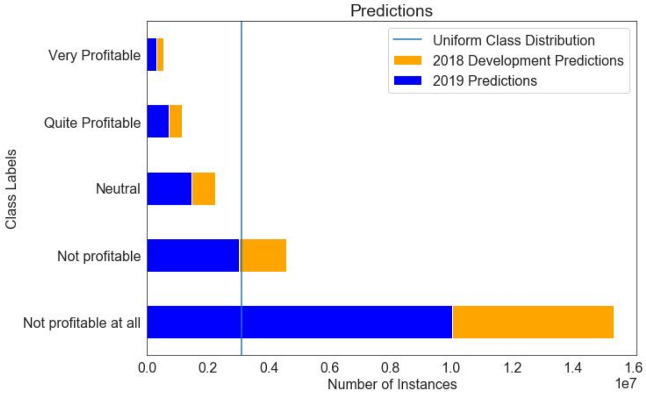


Figure 26: Predicted Class Label Distribution. The Uniform Class Distribution is the number of predictions divided by the number of class labels.

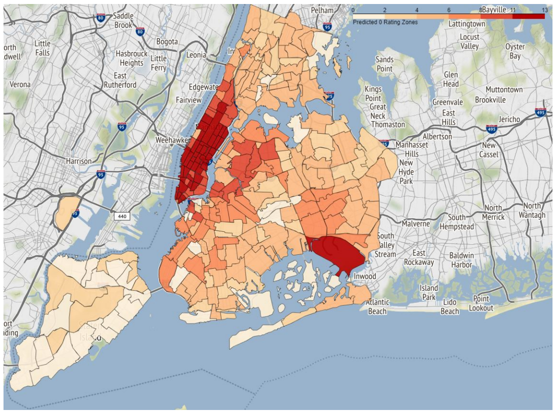


Figure 27: The number predicted outcomes of zones with a 0 rating overall.

Following the hyperparameter tuning, the new Decision Tree was trained and post-pruned from 2669 leaf nodes to a total of 3011 leaf nodes. As such, all leaf node had Gini Impurities lower than 0.4 - a significant improvement when compared to the initial model. It must be noted that the tree is now not interpretable by inspection, and so a Learning Curve (Figure 25) is used to determine the performance and possibility of overfitting.

From observing (Figure 25), the Training to Cross-Validation scores have yet to converge and so there may be signs of slight overfitting to the train data. Nonetheless, (Figure 24) shows strong evidence against overfitting since it seems to generalize well to the development set. Thus, the 2019 predictions will be made with this model.

The 2019 predictions by the model when compared with the 2018 development set seem to quite consistent, following an exponential like relationship with the majority of trips being predicted as "Not Profitable at All". Therefore, we can assume the 2019 predictions to have the best case of 73% accuracy.

The choropleth plots (Figures 27 - 31) seem to contradict the initial findings from the first phase [1] which contended Central NYC / Manhattan **and** the airports to both be profitable. According to the predictions made, the airports are not profitable despite

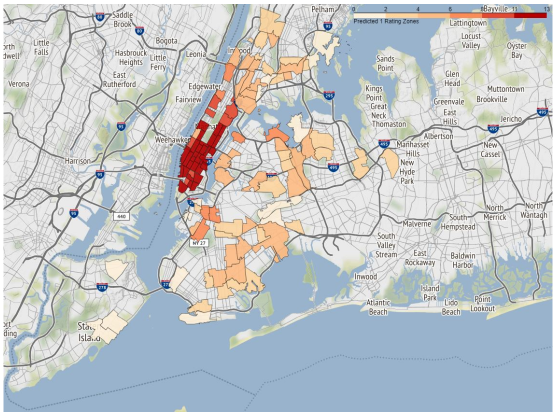


Figure 28: The number predicted outcomes of zones with a 1 rating overall.

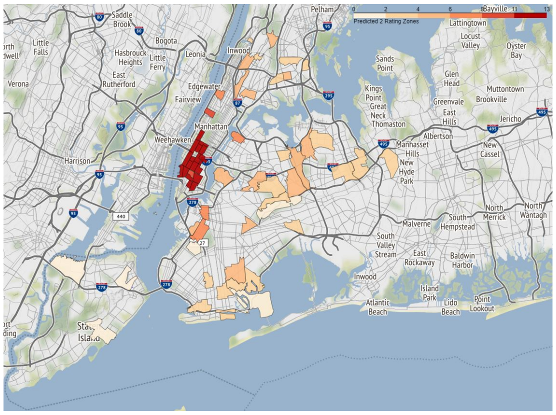


Figure 29: The number predicted outcomes of zones with a 2 rating overall.

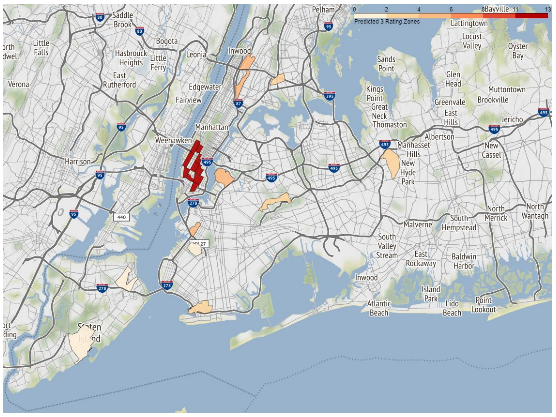


Figure 30: The number predicted outcomes of zones with a 3 rating overall.

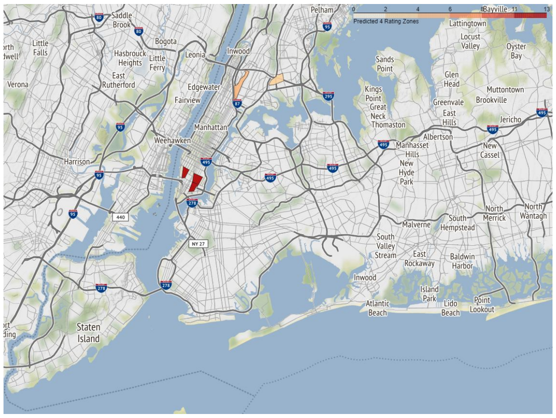


Figure 31: The number predicted outcomes of zones with a 4 rating overall.

the high frequencies. (Figure 28 and 29) also, seem to suggest the surrounding areas **next** to the airports to be more profitable than the airports itself. Furthermore, the 0 rating zones are consistent with the initial findings from the first phase [1] which suggested outer suburbs to not be profitable.

A few unexpected predictions are also present (Figure 30). These zones correspond to Brighton Beach, NYC Health Clinics + Hospitals, Apex Neurology Medical Care + Health Clinics, Green-Wood Cemetery and the cemeteries surrounding Highland Park. Alarmingly one must think, "why is Brighton Beach so popular during months with snowfall???" - this will require further investigation. Reading the Trip Advisor Reviews [8] seems to unlock the secrets behind the unexpected popularity during snowfall:

> "For slavic people... you would have probably better understanding what is oging on around. Because, as soon as you get there, you will realized that you are back to USSR, who lived in it, will understand it. Also. from everywhere you will here either Russian or Ukranian language."

According to locals and visitors, the beach is very popular during the wintertime for those of Ukrainian and Russian backgrounds and is well know for its prime location to view beautiful sunrises. Perhaps this is far fetched, but several reviews suggest that this is the "Russian Winter Experience" in the USA.

However, the main conclusion drawn from the plots deems Central NYC / Manhattan to be the most profitable location for Taxi Drivers. Interestingly, it seems Lower Manhattan is the most saturated and highest demanded location which may be explainable since it is the CBD. Wall Street and the NYC Stock Exchange boast the highest frequency and profit rates, with the surrounding zones in Lower Manhattan yielding similar results.

Finally, according to the DT predictions for 2019:

- Central NYC / Manhattan will be the most profitable and highly demanded areas for 2019 during months with snowfall.
- If a Taxi Driver is not within a close distance to those zones, they can try locally optimal hot-spots which include local shopping districts, hospitals, beaches and recreational centres.
- The first phases' definition of an "Operable Zone" during snow months are consistent with the current findings.

As mentioned in the first phase [1], roads that connect these hot-spots should also be prioritised during heavy snow fall to ensure some form of mobility and basic transportation to be available. Manhattan and Airports should be a high priority in maintaining functional transport routes,

8 Conclusion

The investigations from this analysis found that zones with high demand were deemed most profitable, even if the individual trips were not profitable themselves - a suggestion that contends large quantities of short trips outweighs a single profitable trip. Although the assumptions made in the analysis disrupt the chances in of calculating an accurate estimate of earnings, there is a relationship found - the predictions for months with snowfall in 2019 reveal hot-spots, which are evenly spread out in the outer suburbs.

Since this analysis was conditioned on months with snowfall, a future continuation of this analysis should look at months without snowfall and make a comparison. It is expected to have quite a different distribution of trips and hot-spot locations. Also, the devised rating system was able to fit a Skew Normal Distribution, so a simple goodness-of-fit test to see if the data truly fits the Skew Normal Distribution should be tested. If the null hypothesis is accepted, then a GLM using the Skew Normal Distribution link function given sufficient computational and time resources can be used to accurately predict potential earnings.

Ultimately, the realistic attributes (location, time and weather) seem to be sufficient enough to predict the current zones profitability. Therefore, Taxi Drivers may be able to run this model and decide if they should remain in this zone or move to another.

References

[1] First Phase (Visualization):
https://github.com/akiratwang/NYC-TLC-Making-the-Best-Yellow-Taxi-Driver/blob/master/Phase%201/Visualization.pdf

[2] “TLC Trip Record Data.” About TLC - TLC. Accessed September 7, 2019.
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[3] “Taxi Fare.” Taxi Fare - TLC. Accessed September 7, 2019.
https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.

[4] esm. “Wesm/Feather.” GitHub. Accessed September 7, 2019.
https://github.com/wesm/feather.

[5] “Descriptive Statistics: Skewness and the Mean, Median, and Mode.” OpenStax CNX. Accessed September 7, 2019.
https://cnx.org/contents/bE-w34Vi@9/Descriptive-Statistics-Skewness-and-the-Mean-Median-and-Mode.

[6] “Toll Rates for Bridges and Tunnels.” TollGuru. Accessed September 7, 2019.
https://tollguru.com/toll-info/new-york/toll-rates-for-new-york-bridges-and-tunnels.

[7] “SPSS Tutorials: Pearson Correlation.” LibGuides. Accessed September 7, 2019.
https://libguides.library.kent.edu/SPSS/PearsonCorr.

[8] “Brighton Beach (Brooklyn): UPDATED 2019 All You Need to Know Before You Go (with PHOTOS).” TripAdvisor. Accessed September 7, 2019.
https://www.tripadvisor.com.au/Attraction_Review-g60827-d116390-Reviews-Brighton_Beach-Brooklyn_New_York.html.

[9] “Use Jupyter Notebook Remotely.” Use Jupyter notebook remotely - pytraj 2.0.2.dev0 documentation. Accessed September 7, 2019.
https://amber-md.github.io/pytraj/latest/tutorials/remote_jupyter_notebook.

[10] “Scikit-Learn.” scikit. Accessed September 7, 2019.
https://scikit-learn.org/stable/index.html.