Exploring existence of relationships between extreme winter weather using machine learning models

MAST 30034

Insert Student Yujie Li
Student ID: 1174055
Github repo with commit

August 31, 2022

1 Introduction:

In recent years, extreme weather has caused severe consequences worldwide, especially traffic issues. New York also experiences lots of weather-induced traffic stress during wintertime, with snowstorms blanketing almost all streets in New York[1] and bringing ground commutes to a standstill. Therefore, the overwhelmed transit system needs to be carefully analysed for both taxi companies and public transport systems to make foresightful decisions. As the extreme destructive weather occurs in the city mostly during winter, it is crucial to ensure that all taxi drivers are paid fairly and deserve compensation for working in such an extreme environment. Hence, it is essential to uncover whether there is the underlying relationship between weather and the income of taxi drivers during winter.

In this study, **we aim to explore if there exists linear relationships between extreme winter weather and taxi drivers' income using machine learning models**.

Instead of applying the traditional way to analyse the potential relationship by performing an F-test on each parameter or t-tests on differences in means of different groups, this research will look to more empirical evidence on the relationship between weather data and yellow taxi drivers' income by analysing the fit and generalisability of machine learning models.

To identify the income of drivers, this study will create a new metric called income rate by the ratio of fees paid to the driver and the time of travelling. Besides, preprocessing will be performed before the above feature engineering to remove potential outliers and human error. Feature selection will also be established to select the most significant features and avoid overfitting. Then, we will employ linear regression and random forest regression to model and analyse the relationship between weather and income rate to give recommendations on the relationship between extreme weather and income rate for taxi companies and public transit systems.

2 Dataset:

The primary dataset used in this study is the yellow taxi data published by the NYC Taxi Limousine Commission (NYCTLC). This dataset records trip statistics for all New York City taxi and For-Hire Vehicles (FHV). The dataset includes trip features such as trip time, distance, the amount paid and tips (only in card payment). To investigate the taxi drivers' income performance in winter, this research

only comprises data from 2016 December and 2017-2018 Winter (January, February, November and December). Hence, the primarily used dataset in this research will contain 81,725,226 records of taxi trips with 19 features.

The external dataset includes in this study is from Visual Crossing Corporation[2], starting from 2016 to 2019. The weather features from external datasets will add extra predictors for predicting the income of taxi drivers during winter.

The preprocessing and aggregation method will be further explained in later sections.

3 Preprocessing:

3.1 Data cleaning

After downloading raw taxi data in parquet form, we first choose nine features related to our study based on intuition. ['PULocationID', 'DOLocationID', 'tpe_pickup_datetime', 'tpep_dropoff_datetime', 'trip_distance', 'payment_type', 'fare_amount','tip_amount', 'total_amount'] We clean data that may incur a human error when recording from different dimensions for taxi data each month.

1. Removing data out of month time range: We remove data with pick-up and drop-off time that is not in the time range of the data month.
2. Removing data with negative trip distance and tip amounts: From (scatterplot), it is demonstrated that some trip distance is less than zero, which contradicts the common understanding of trip distance. Hence, we remove these data points. Similarly, we eliminate data with negative tip amounts.
3. Removing data with fare amounts less than the initial charge: From the NYCTLC website[3], the initial charge for each yellow taxi trip is $2.5. Hence, records with a fare amount less than $2.5 should be removed to stick with the payment rule of yellow taxi trips.
4. Removing data with speed over legal limit: Besides, by the law of New York, the limited legal speed in the city is 65 mph (miles per hour)[4]. Hence, we will also take out records with an average speed of over 70 mph since they are not very realistic. These data are not meaningful in studying the majority trends of yellow taxi drivers' travelling time

3.2 Outlier detection

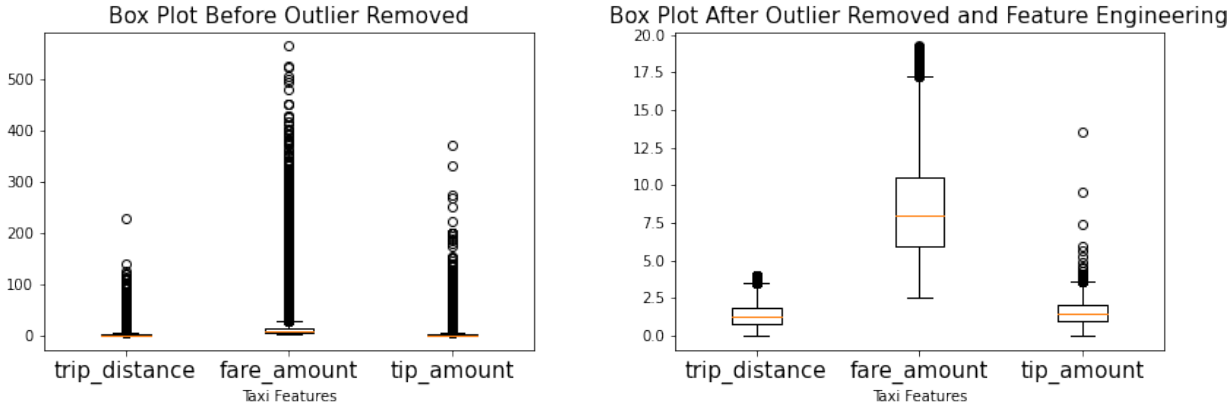
After removing the data conflicting common sense, many unrealistic records remain in the dataset, as shown in the following boxplots (see Figure 1a). It is assumed that the taxi data will have similar performance across all months in this study. Hence, quantiles from the December 2016 dataset is used as an example to remove data ranges that are not realistic. We will examine extreme outliers shown in boxplots and eliminate them if their features do not seem reasonable.

3.3 Using IQR to further eliminate

After detecting unreasonable data points, the data distribution still seems highly positively skewed. Therefore, we need to use IQR to further manipulate feature distribution in the taxi dataset. Since a large number of outliers are filled in the upper range, it is more important to restrict extreme large data. Hence, we employ a range of $(Q1 - 1.5IQR, Q3 + 0.5IQR)$ clean statistical outliers instead of traditional $Q3 + 1.5IQR$ from the upper fence.

3.4 Feature engineering

1. **Filling tip amount:** Since we are investigating the total income of taxi drivers for each trip, it is necessary to fill in the tip amounts not given in the original data. The histogram below illustrates that only payment method 1(credit card) and method 2(cash) play a key role among all six payment methods. Hence, we can first eliminate trips that do not include payment methods 1 and 2. According to the original dataset, tip amounts made by credit cards are automatically populated; however, tips for cash payments are not included in the table. Hence we need to impute in tip amounts for rows with cash payments. It is assumed that tip amounts distribution for both payment methods will be identical. Thus, we can deduce the estimated tip amounts for cash payment trips by calculating the mean tip rate of credit card payment trips and multiplying it to cash fare amounts. To achieve this, we first need to compute the average tip rate of payments made by credit card from the recordings by dividing the tip and fare amounts beyond the $2.50 initial charge for each credit card trip and then taking its mean. Finally, we complete the cash tip amount by multiplying the mean credit card tip rate and the fare amounts made by each cash payment if the payment method is by cash. We use the sum of tips (now both cash and card) and fare amounts for each trip to replace the previous 'total amount'.
2. **Calculate income rate metric:** To explore the possible effect between weather and taxi drivers' income, we need a metric that measures the driver's income for each trip. As claimed in the yellow trips data dictionary on NYCTLC, the fare amount is the time-and-distance calculated fare. It is reasonable to use travelling time to deduce the income rate of every single trip. Therefore, this study will employ an income rate metric, the ratio of the total payment and travelling time for each trip, to indicate how much the driver receives for every mile travelled in the trip (after preprocessed data see Figure 1b).



(a) Boxplot of 2016.12 taxi features Before Outlier Removing

(b) Boxplot of 2016.12 taxi features After Outlier Removing

3.5 Aggregation

Since the granularity of the external dataset is daily, the cleaned taxi dataset needs to be aggregated by date to merge with the weather data. Therefore, we first aggregate the taxi data by pick-up location (PULocationID) and then by the date of the trip. After aggregation, we merge it with the external dataset on the date column. Since too many features are included in the external dataset, overfit problem may occur due to the curse of dimensionality. As a result, training accuracy will be excessively higher than validation and testing accuracy. To avoid the curse of dimensionality, we first performed intuitively feature selection from the external dataset based on intuition relating to the

winter weather and its potential impact on taxi driving. Later, we also conducted an F-test feature selection to further reduce the dimension (which will be specified in later sections). After data cleaning and aggregation by PULocationID and date, there are 52,822 records left for modelling.

3.6 Fill in Null Values

After merging taxi and weather datasets, it is noticed that null values exist in the merged dataset. As all null values are shown in the "windgust" feature and the minimum of this feature apart from Null values is 25, it is assumed that the existence of null values is due to zero wind gust on such day. Therefore, we decided to fill all rows with NaN wind gust as zero.

3.7 Train Test Split

We split the dataset by holding 60% for training data chronologically since we need to use historical data as the training set to predict validation and testing sets in the future. Then equally divide the remaining as the validation and testing set, using training data to fit the model and validation data to choose the best hyperparameter combination. We randomised the order of data when splitting validation and testing sets instead of chronological splits, as the validation and test set should be similar in characteristics by definition.

3.8 Standardisation

Since our computational resources are limited for this research, it is critical to standardise all numerical predictors and thus save storage size and create a faster tuning process in the later modelling section.

4 Preliminary Analysis:

4.1 Baseline

To measure and compare models' performance, we need to find a benchmark of regression models for the dataset we are predicting. The formula for the linear regression line is $Y = \alpha + \beta_1 X + \epsilon$. The null model will follow the null hypothesis (H0), where $\beta_1 = 0$. Therefore, it will produce a regression line which is just the sample mean of y without any predictors involved, $Y = \alpha + \epsilon$. By definition, the R_square for the testing dataset is zero. This is because nothing is included in the regression but the mean of the response variable. The RMSE (root mean square error)

$$RMSE = \sqrt{\frac{\sum_{n=N}^{i=1}(x_i - \hat{x}_i)^2}{N}}$$

measures the absolute fit of the model to the data, indicating how concentrated the data is around the regression line across the sample mean of income rate. A low RMSE suggests a better fit and is a good measure for determining the accuracy of the model's predictions. The RMSE for the training set is more significant than the validation and testing set, indicating that no overfit problem occurs in the null model (see Figure 2).

4.2 F-test for feature selection

After intuitively selecting factors, we implement statistical feature selection based on F-test to decide factors that may produce an effect when predicting drivers' income in winter. This process of reducing the number of input variables reduces training time and, in some cases, avoids fitting noise and

```
r2_score for training(dummy mean): 0.000
r2_score for testing(dummy mean): -0.003
Mean squared error (dummy): 6.277
Root mean squared error for training(dummy): 4.729
Root mean squared error for validation(dummy): 3.597
Root mean squared error for testing(dummy): 2.505
```

Figure 2: Performance of Null Model

improves the model's accuracy. We use SelectKBest + F-regression test from Scikitlearn to get optimal eight features across factors of four factors from the taxi dataset and thirteen factors from the weather dataset. The F-regression test measures the statistical significance of one parameter in predicting the response variable. The larger the f-score is, the more meaningful the predictor will be and will be included in the linear regression model later.

The regression result indicates that four features from the primary dataset are all significantly associated with income rate. However, only four features from weather data show significance with income rate, and none of them is as significant as the features from the primary dataset. This may be due to the relatively weak relationship between weather and the income of drivers. However, the largest proportion of drivers' income is from the fare amount charged for the trip, which is calculated both time and distance based. Hence, the time and trip distance features should highly correlate with income (see Table 1).

total_adjusted_amount	trip_distance	fare_amount	tip_amount
3230.172	683.986	676.351	1164.063
precipcover	humidity	cloudcover	visibility
1.776	2.380	2.989	1.398

Table 1: SKB

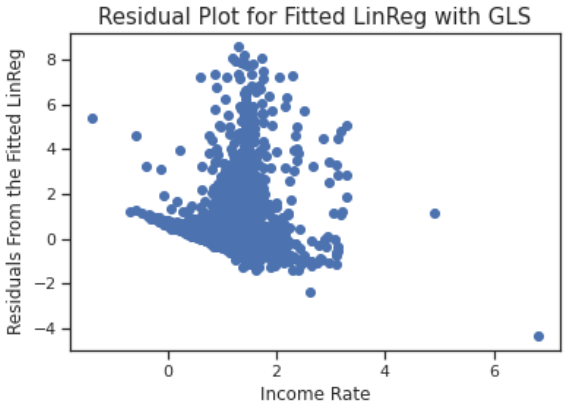
Besides, although the factors we hypothesised that should be highly correlated with income rate, such as "snow" and "temperature" are not included in the optimal eight features. The factors selected by the f-test are still intuitively reasonable. For example, the "precip-cover" factor measures the percentage of time during the reporting window when the precipitation occurred[5]. It is hypothesised that when precipitation coverage is high, drivers have to drive slowly and carefully, where the travelling time will be longer than expected. The percentage of traffic accidents and risk of driving will also be significantly increased; hence, it is also reasonable for drivers to be paid more than usual. However, for factors such as "snow", it is very hard to measure the snowfall amount for daily granularity since snowy days may not occur that frequently in a month. Thus, the influence of such a factor may not be as significant as other measurements that will happen every day.

5 Model:

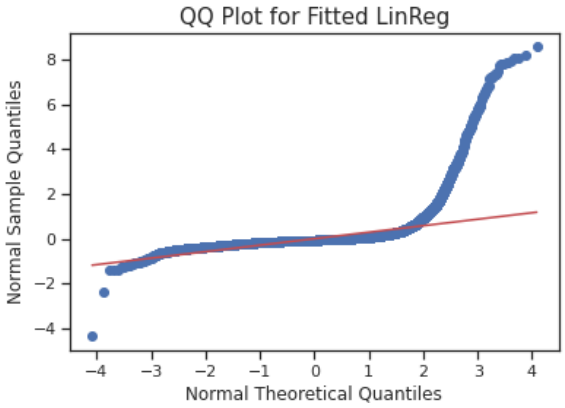
5.1 Linear Regression (LR)

One statistical model we will carry out in this research is Linear Regression. Linear regression is a model that fits a straight line between predictors and the response variable and measures their relationship. The formula of linear regression is $Y = \alpha + \beta_1 X + \epsilon$. In this research, the predictors will be the eight features selected from the feature selection with F-test, and the response will be

the income rate of yellow taxi drivers. We first sketch the scatterplot by pairs to detect if there are any present linear relationships. The scatterplot illustrates that the most extreme values are shown when the income rate is larger than ten. These data highly affect the distribution of scatterplots, and they may be outliers that are not significant to the relationship we aim to learn. Hence, we decide to truncate the data to less than ten for income rate to see the scatterplot trend more clearly.



(a) Residual Plot for Fitted Linear Regression



(b) QQ Plot for Fitted Linear Regression

There are essential assumptions for this regression method to be suitable for fitting. The first assumption is homoskedasticity in the dataset, which means the residuals should not change dramatically across all values of dependent variables. Figure 3a, the residual plot of the training dataset, illustrates that the homoskedasticity assumption is not valid in this dataset. To fix it, we will implement generalized least squares (GLS) instead of ordinary least squares (OLS). Unlike OLS, which assumes all variances in the dataset have the same mean,GLS will use weighted least squares, where larger variances have a smaller weight [6]. Besides, linear regression also assumes data's normality, meaning residuals will follow a normal distribution. From the QQ-plot of the data (see Figure 3b, although most left tails follow the normal distribution, the right tail seems dramatically conflict with the normality assumption. However, since the right tails represent unrealistic large income rates to our intuitive understanding, we can still interpret them as unmeaningful outliers that will not affect our distribution. Hence, we will assume this data satisfy the normality assumption of linear regression.

After checking the assumption of linear regression, we tried to transform the data to check if the data could be better fitted into the linear model. However, after applying log transformation to "total_adjusted_amount", "trip_distance", "tip_amount" and "fare_amount", it is found that the R-squared value is decreasing from 0.182 to 0.088. This is probably because data trends are hard to visualise simply from scatterplots, and the log transformation of predictors causes the model's decrease in linearity. Hence, we decide to retain the original linear regression with GLS. The formula of the final regression line is

$$\begin{aligned} income\_rate = {} & 1.122 + 0.703 * total\_adjusted\_value + 0.262 * trip\_distance - 0.934 * fare\_amount \\ & - 0.133 * tip\_amount + 0.001 * precipcover + 0.015 * humidity - 0.003 * cloudcover \\ & + 0.020 * visibility \end{aligned}$$

The regression formula indicates that the precipitation coverage, humidity and visibility will positively affect the drivers' income. From the t_test result in the following figure, it is also depicted that humidity and visibility significantly affect drivers' income since they are less than 0.05. The RMSE for the validation set is 3.152 and 2.182 for the testing set, which are both higher than the RMSE for the training set (0.469). This may also be due to the overfitting of the regression line. The R_squared

value for the validation set is 0.232, and the testing set has a R_sqauare value is 0.239, which are both higher than the training set.

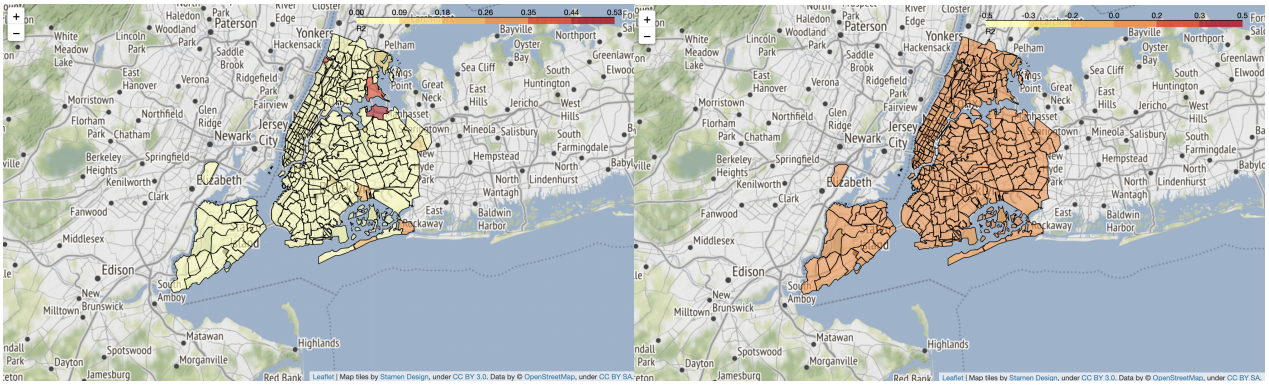
5.2 RFR

Another efficient machine learning algorithm we will implement in this study is Random Forest Regression, an ensemble learner that uses combinations of multiple random decision trees. This algorithm will employ the result from each tree with different input data samples. Different features will be selected at each node, and the trees will be built independently without interaction. Each end node gets regression performed on it, and the result from each tree will eventually be averaged to produce the final prediction of Random Forest as a stacking algorithm. There is no need to feature select before performing RFR since the algorithm will feature select by its embedded algorithm as a class of random trees.

After tuning with grid search, the optimal hyperparameter combination will be 'ccp_alpha': 0.001, 'max_depth': 30, 'max_features': 0.75, 'max_samples': 0.5, 'n_estimators': 60. The RMSE for the training dataset is 1.795, while the validation set has an RMSE of 31.622 and 31.586 for the testing set. The massive difference between training and the other two sets is probably due to overfitting and the curse of dimensionality. As the model becomes more complex with multiple regressors, it works on captures most of the training data pattern and reduces its RMSE to a small number. Thus, the model tends to generalise the training data trend rather than the true properties of the dataset. By this, noises in the training data may be treated as useful information and be learnt, and the model will exhibit overfit. When the dataset is replaced by validation and testing data, the model can no longer fit the regression learned from training data onto the new dataset well; hence, a large RMSE exists.

5.3 Comparison

Comparing the testing RMSE from both LR and RFR to the Baseline model, it is discovered that only LR has a lower testing RMSE than the baseline model. This reveals that the performance of LR is meaningful as a model since it has a better absolute fit to the data. In contrast, the RFR model has a higher testing RMSE than Baseline, indicating a relatively poor performance in exploring the relationship between drivers' income and weather data. To investigate more on the performance of R_square in the model conducted in this empirical analysis. We introduce two heatmaps based on each locationID with R_sq as the underlying value. It was discovered that the R_square for most locations is less than zero, where only 14 locations give a positive R_square number in the LR model, and none of the 263 locations are positive in the RFR model. A negative R_square means the model is giving a higher sum of square error than the individual null models (note: all negative R_square values are replaced by zero for better data visualisation). Hence, the model performance in this research is not stable enough, where models perform worse than taking the mean of the response when aggregating by location. The two maps depict that although both models do not constantly perform well when specifying locations in New York, Linear Regression maintains a relatively better result. This difference is probably due to many noises in the data are fitted into both models, and RFR is more sensitive to noises than LR. The tuned RFR model will fit 60 trees and employ significantly more features than LR will use; hence, when the data are noisy or not highly correlated to independent variables, RFR will overfit.



(a) R^2 Map for LR (b) R^2 Map for RFR

6 Recommendation

In this study, we are trying to figure out the empirical analysis by building machine learning models instead of just applying statistical tests. However, from the result of model performance and their comparison, we found that the models we implement do not exhibit a significant effect on finding the relationship, and hence currently, bad weather is not having a significant effect on drivers' income rate. There are several ways in which we can build more precise models in the future, one of them being adding more data to the analysis. For example, if the weather data's granularity can be refined, such as hourly or data measured in different locations. We can also find the best fit LR for each location by modelling a single LR for each locationID in the future, to better investigate for relationships. However, this will require a large amount of computational resources and enough external datasets to support the investigation.

From the results of this research, it is recommended that taxi companies consider compensating drivers during bad weather by increasing the fare amount per distance, and also adding flat out bonuses for the extra risk they are bearing, as travelling in extreme weather conditions increases the risk of accidents. Actuarial should be hired to carefully analyse the risk associated and derive the bonus amount.

References

[1] Jeff Berardelli. *Snowstorm causes havoc in NYC – what went wrong?* https://www.cbsnews.com/news/snowstorm-nyc-traffic-port-authority-forecast-what-went-wrong/. Accessed: 2022-08-10.

[2] Visual Corporation. *Weather Data Services — Visual Crossing Corporation*. https://www.visualcrossing.com/weather/weather-data-services. Accessed: 2022-08-05.

[3] Visual Corporation. *Taxi Fare — Visual Crossing Corporation*. https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page. Accessed: 2022-08-07.

[4] Dave Werner. *Speed limit 85 MPH*. https://www.adirondackdailyenterprise.com/opinion/columns/safety-on-the-roads-by-dave-werner/2019/07/speed-limit-85-mph/. Accessed: 2022-08-06.

[5] Sam Helwig. *Precipitation Coverage is the Most Important Weather Metric You don't Know — Visual Crossing Corporation*. https://www.visualcrossing.com/resources/blog/what-is-precipitation-coverage/. Accessed: 2022-08-06.

[6] Unknown. *Heteroskedasticity and GLS*. https://www.reed.edu/economics/parker/312/notes/Notes7.pdf. Accessed: 2022-08-06.