WEATHER and UBER RIDES in NEW YORK

(2014 & 2015)

An Analytics & Machine Learning Based Report

Author: Raghuraman Srinivasan

1. Problem

To Test the Hypothesis "Weather is a predictor of volume and duration of uber rides in New York" using Analytics and Modeling.

2. Data

The data used in this hypothesis testing is sourced from NYC Uber trips and NYC historical weather datasets available from the following links.

Data Link:

- Uber trip data in NYC - https://github.com/fivethirtyeight/uber-tlc-foil-response
- Historical weather data for NYC - https://www.kaggle.com/selfishgene/historical-hourly-weather-data

Format:

- Uber trip data in NYC
  The data source contains 7 CSV files. Six of the CSV contains Uber trip data from April 2014 to September 2014, and one CSV file contains Uber trip data from January 2015 to June 2015.

- Historical weather data in NYC
  The data source contains 6 CSV files. All of them contain different weather information such as temperature, humidity, pressure, wind speed, wind direction and weather description of multiple cities.

Preparation:

- Datasets are loaded into the Jupyter Notebook's Python kernel and all the data operations are performed using the "*pandas*" library.

- All individual CSV files in both the datasets are initially merged while loading with only the potentially useful variables and then the variable names are modified accordingly. The historical weather data is filtered with only the entries of the New York city.

- The Uber trip data is named as "*nyc_uber*" and the historical weather data is named as "*nyc_weather*".

- Data exploration has identified the fields with incorrect data types on both the data sets. All such fields are defined with appropriate data types.

- The "*nyc_uber"* data is aggregated by the number of hourly "*Rides*" and is grouped by the "*Base*" of the booking.

- The "*nyc_uber*" data is also aggregated by "*Month*" across both the years 2014 & 2015 separately to understand the standout months.

- Data exploration in "*nyc_weather*" has identified **38 missing values out of the 8690 values** in the "*Humidity*" column.

- The values in the "*Temperature*" column in "*nyc_weather*" data are changed from Kelvin to Celsius using the following formula

Temperature in Celsius = Temperature in Kelvin – 273

- The weather "*Description*" column in the "*nyc_weather*" contains multiple categories that closely convey the same weather condition. All such categories are grouped together and the total number of categories has been **reduced from 25 to 12 categories**.

- The "*nyc_uber*" and the "*nyc_weather*" data are saved as CSV files in the respective names.

- The "*nyc_uber*" and the "*nyc_weather*" data are then inner joined together based on their "*DateTime*" columns and the resulting data frame is named as "*nyc_uber_weather*".

- The joined "*nyc_uber_weather*" data has a total of 189 missing values out of 48824 values in the "*Humidity*" column.

- Date and time related features such as "*Year*", "*Month*", "*Day*", "*Weekday*", "*Hour*" are extracted from the "*nyc_uber_weather*" data to explore the cyclic data patters.

- Data dictionary of the "*nyc_uber_weather*" data (*Table 1*)

#	Column	Description
1	DateTime	Timestamp of a Uber ride
2	Base	Base of the Uber
3	Rides	Number of Uber rides
4	Description	Condition of the weather
5	Temperature	Temperature in Celsius
6	Humidity	Humidity level
7	Pressure	Pressure level
8	WindSpeed	Speed of wind in Km/Hr
9	WindDirection	Direction of wind
10	Year	Year of Uber ride
11	Month	Month of Uber ride
12	Day	Day in Month of Uber ride
13	Weekday	Weekday of Uber ride
14	Hour	Hour of Uber ride

Table 1. Data dictionary

3. Exploratory Data Analysis

Statistical summary:

Statistical summary of the "*nyc_uber_weather*" data (*Table 2*) conveys the distribution of the columns using statistical metrics. The following points are inferred from "*Table 2*".

- The "Rides" column has a standard deviation value almost twice the mean value indicating high variance between values and abnormal distribution of the data.
- The "*Temperature*" column has the standard deviation value almost close to the mean value indicating the exponential highs and lows in "*Temperature*" values.
- Humidity percentage stays on the higher side majority of the days with mean value of 68.57.
- There is not much significant inferences from the remaining columns.

	Rides	Temperature	Humidity	Pressure	WindSpeed	WindDirection	Year	Month	Day	Hour
count	48824.000000	48824.000000	48635.000000	48824.000000	48824.000000	48824.000000	48824.000000	48824.000000	48824.000000	48824.000000
mean	111.598620	12.593225	68.576128	1019.506902	2.622276	203.332726	2014.554686	4.894990	15.568839	11.542889
std	229.645724	11.454197	18.999282	9.223876	1.737264	100.811156	0.497006	2.239791	8.714739	6.913471
min	1.000000	-22.226000	10.000000	968.000000	0.000000	0.000000	2014.000000	1.000000	1.000000	0.000000
25%	40.000000	4.147500	54.000000	1013.000000	1.000000	133.000000	2014.000000	3.000000	8.000000	6.000000
50%	60.000000	15.540000	70.000000	1019.000000	2.000000	214.000000	2015.000000	5.000000	16.000000	12.000000
75%	60.000000	21.370000	84.000000	1024.000000	3.000000	290.000000	2015.000000	6.000000	23.000000	18.000000
max	2346.000000	33.820000	100.000000	1052.000000	13.000000	360.000000	2015.000000	9.000000	31.000000	23.000000

Table 2. Statistical Summary of the Data

Correlation Plot

Fig 1. Correlation Chart

Based on the correlation plot (*Fig 1*) the columns "Rides", "Temperature", "Humidity", "Year", "Month" shows strong correlation values with other columns. So, starting with analytics around these columns will take the solution in the right direction.

Distribution of Rides Across the Timeline:

The distribution of the "*Rides*" representation in (*Fig 2*) infers the following

- The number of hourly rides are usually below 100. Only in the month of January and February 2015 shows an abnormal increase in the number of hourly rides reaching over 2000.
- The "*Base*" *B20764* shows a steep increase in hourly rides followed by *B20617* . Surprisingly these bases have been with less hourly rides compared to the other bases.
- Since the months January & February marks the winter season in New York, the distribution of "*Rides*" based on weather condition and indicating values such as temperature, humidity might be a worthy look.

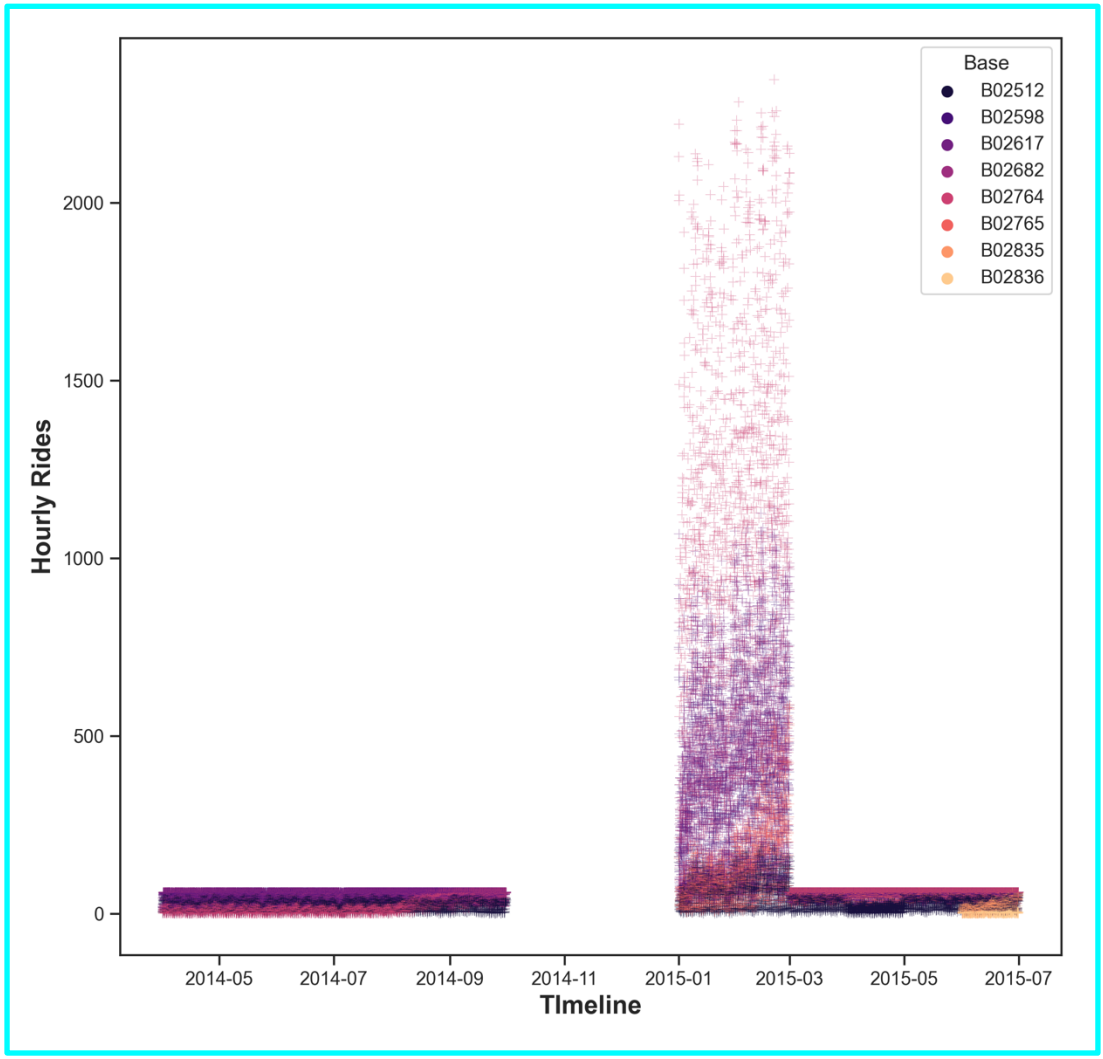

Fig 2. Distribution of Hourly Rides across Timeline

Distribution of Rides as a variation of Temperature:

The distribution of rides based on temperature changes (*Fig 3*), the variation of temperature across timeline (*Fig 4*), and the distribution of rides across weather conditions, temperature and humidity (*Fig 5, Fig 6*) infers the following

- The number of hourly rides shows a sudden increase in number as the temperature dips below 10 degree Celsius and increases exponentially after the temperature falls below 0 degree Celsius.

- The temperature predominantly goes below 0 degree Celsius during the months of January and February.

- Also, the increase in Uber rides is moderate on rainy conditions, while its high on snowy conditions as that relates to dip in temperature below zero as the humidity levels stay well over 60% when the number of rides is high.

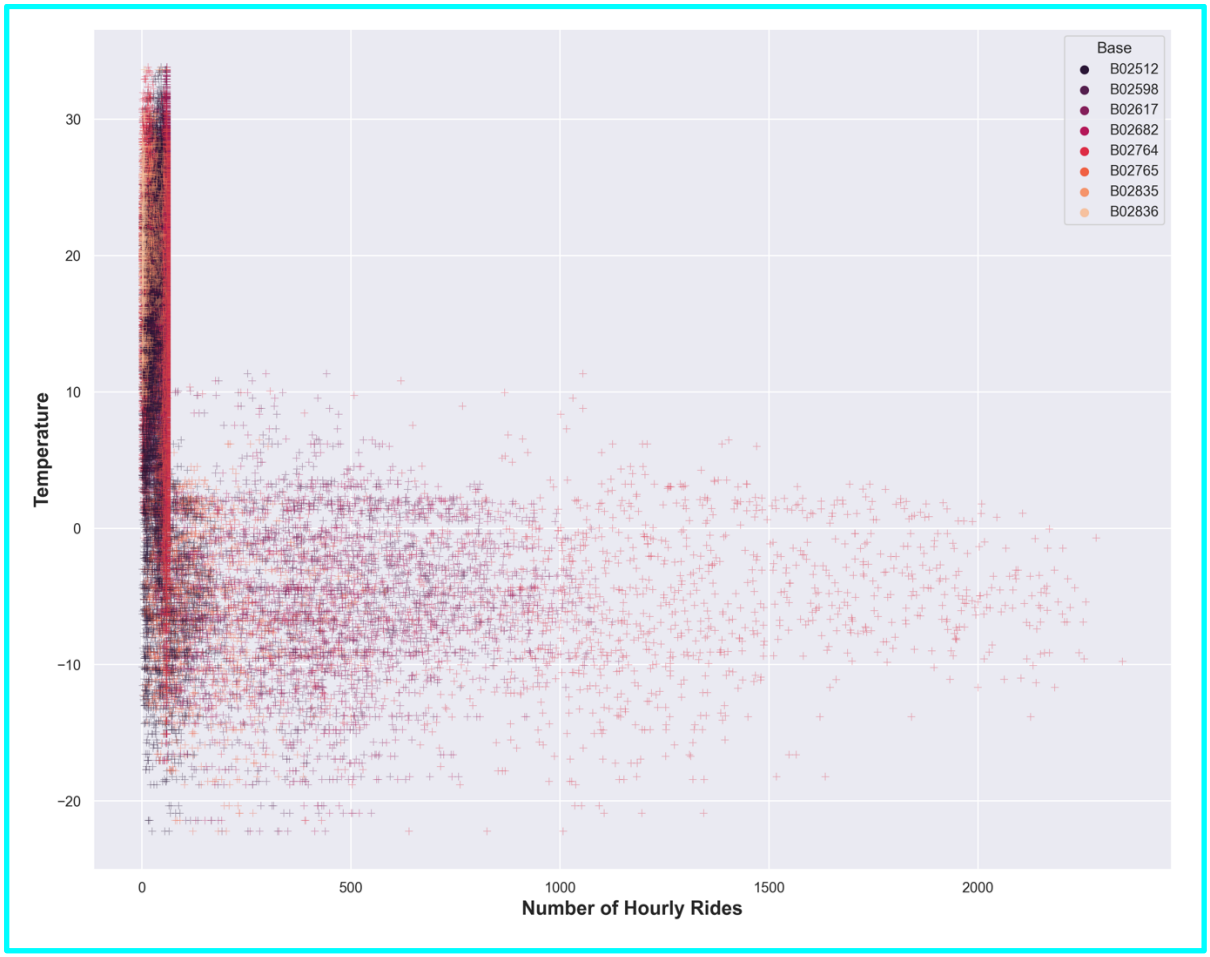

Fig 3. Distribution of Hourly Rides as a variation of Temperature

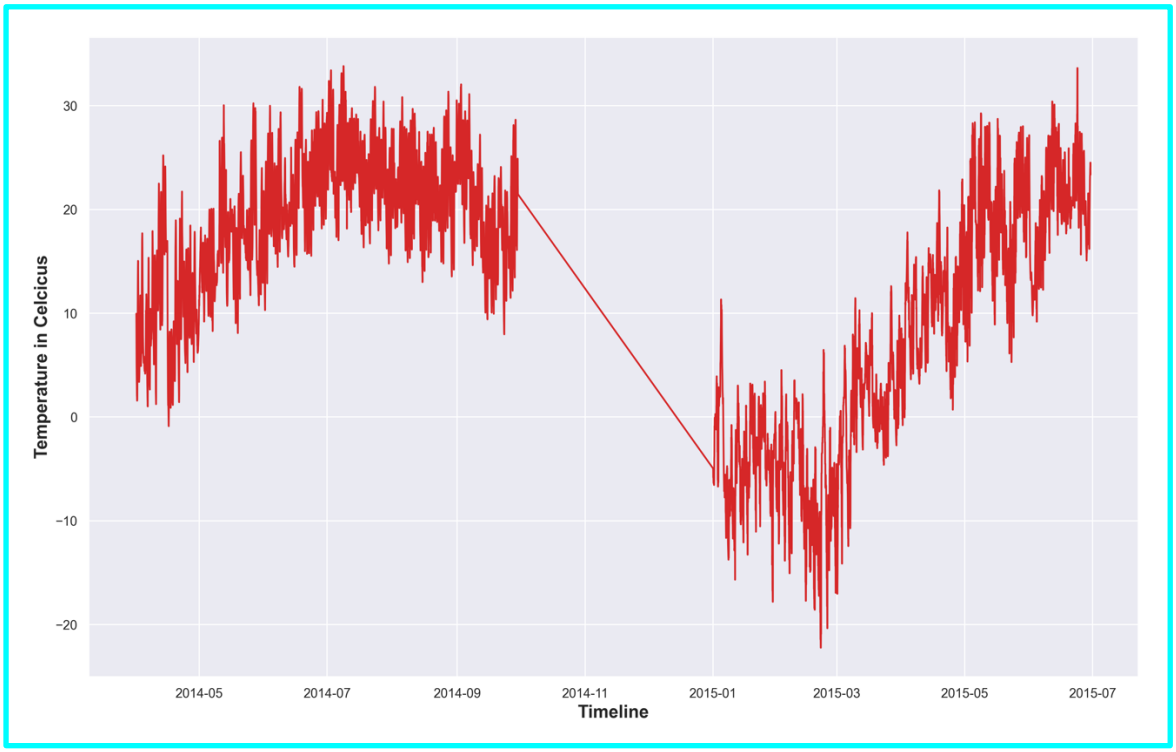

Fig 4. Variation of Temperature across the Timeline

Fig 5. Distribution of Hourly Rides across Weather Conditions

Fig 6. Distribution of Hourly Rides across Temperature and Humidity

4. Modeling

According to the hypothesis, "*Rides*" will be the target variable and the weather will be predictor variables. From the analysis, apart from weather variables such as temperature, humidity and weather conditions, it is seen that the date & time variables along with the base of the Uber also influences the number of "Rides".

Hence, along with weather variables the date time variables and base will be included as predictor variable and infer how well the weather variables influence the "*Rides*" in comparison to the other variables.

The "*Sklearn*" library regression models - Linear Regressor, SGD Regressor and Random Forest Tree Regressor will be implemented due to the following traits.

- **Linear Regressor –** Highly desirable model as it is fast on large training data, it establishes a linear relationship which helps in accurate interpretation of the results and the predicted values can extrapolate which makes it useful in varying unseen data.

- **SGD Regressor:** It is considered as the data is abnormal and contains outliers. The huber loss function in SGD Regressor helps with data containing outliers.

- **Random Forest Regressor:** It is considered as it is robust on the data and can generate highly accurate results. But on the downside, it is time consuming, interpretability is not accurate as there is an element of black box in Random Forest models and it cannot extrapolate outside the seen values which makes it not so useful in varying unseen data.

All the models are pipelined with the following pre-process steps

- One Hot Encoding of categorical variables.
- KNN imputation to impute the missing values in the Humidity column.
- Standard Scaler to scale values as there is an exponential variation observed in certain columns .

5. Results

Model Metrics

Model	Train Accuracy	Test Accuracy	RMSE	R Squared
Linear Regressor	0.3380185	0.354791921	184.1899402664	0.35479192164
SGD Regressor	-0.01691748	-0.021879853	231.8013321281	0.02187985379
Random Forest Regressor	0.986600640	0.9657057678	42.46455237456	0.96570576786

Table 3. Model Metrics

- From the result of model metrics (*Table 3*), it can be seen that the random forest regressor model has performed better compared to the linear regressor model. Meanwhile the SGD Regressor has performed poorly.

- From the above observations, it is worth only looking at the interpretation of the linear regressor model and the random forest regressor model.

- The regression fit of predictions on the test data on the two considered models is shown in (*Fig 7, Fig 8*).

- The random forest has predicted the "*Rides*" perfectly and the linear regression fails to predict the values which are abnormally high.

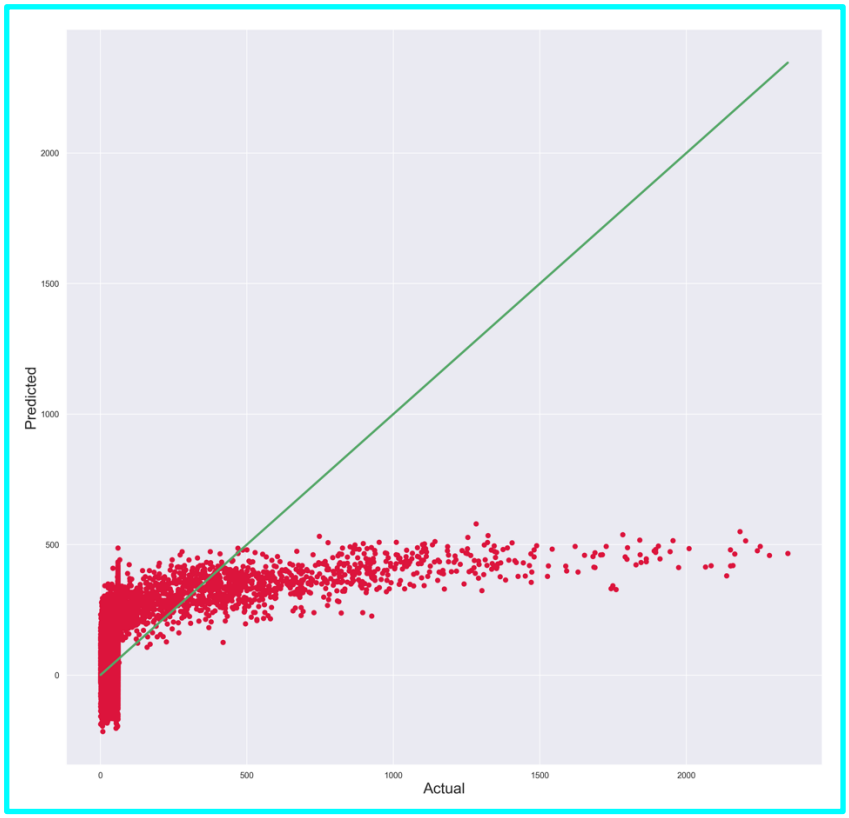

Fig 7. Fit of Linear Regressor Model

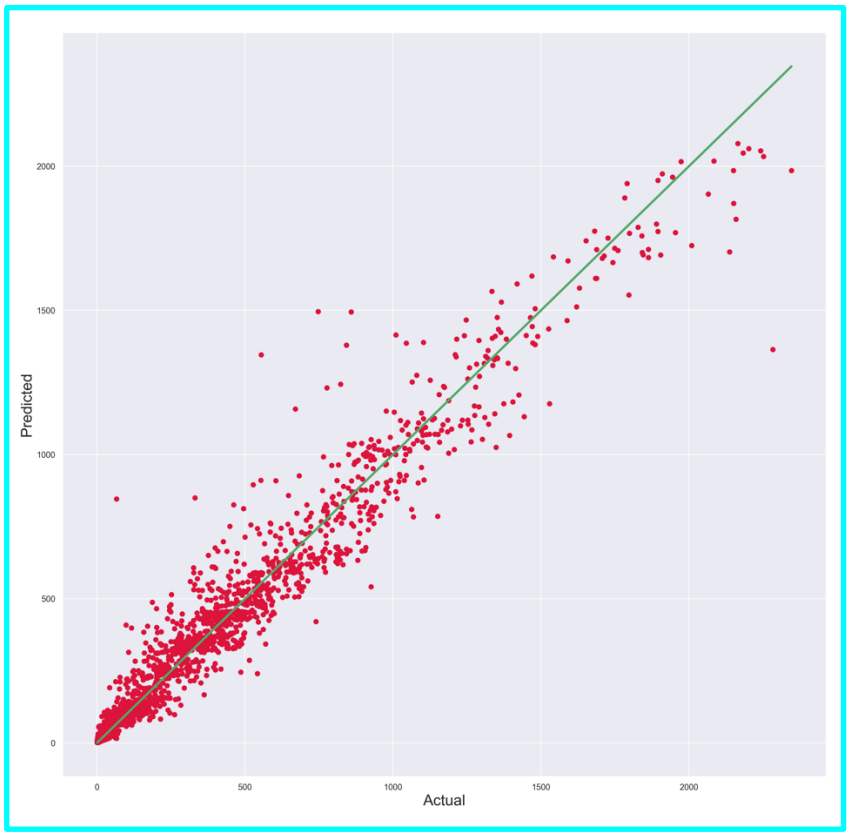

Fig 8. Fit of Random Forest Regressor Model

Feature Importance:

While the random forest regressor has performed better than linear regressor model, the feature importance chart of the random forest regressor is generally inaccurate to its Blackbox nature. The feature importance chart of the linear regressor (*Fig 9*) will be considered based on the following observations and conclusions

- The linear regressor model is interpretable and feature importance chart is accurate.
- Though the accuracy of linear regressor is only 35% on the test data, the RMSE value of 184.2 is less compared to the standard deviation of the "Rides" variable which is 229.6.

On interpretation, it can be seen that the most valuable predictor is the "Temperature" column. The pressure and humidity also have influenced the prediction quite well. Also, all the individual weather conditions in the "Description" columns have influenced the prediction. Combinedly, the weather variables prove to be useful predictors.

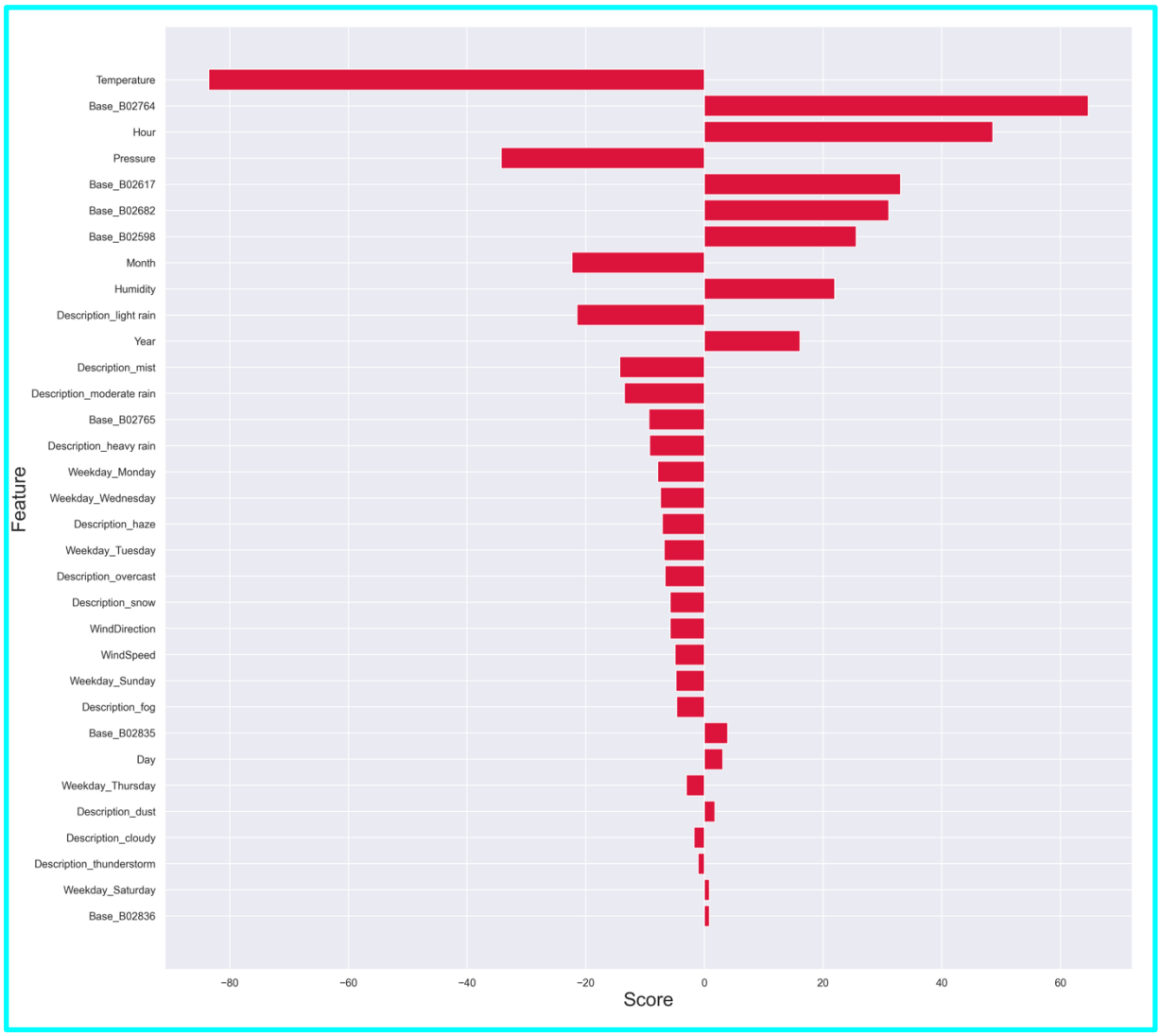

Fig 9. Feature Importance of Linear Regressor Model

6. Conclusion

From the statistical analysis, distribution plots and feature importance charts it is evident that the lowering of temperature (snow & rainfall), humidity (rainfall), pressure (mist, haze, fog) have all influenced towards the prediction of the volume of Uber rides which conclusively proves the hypothesis that there is a influence of weather variables in predicting the volume & duration of the Uber rides.