COURSE PROJECT

IE 7280 STATISTICAL METHODS IN ENGINEERING

TOPIC: NYC Uber pickups with weather and holidays using Multivariate analysis

Northeastern University

TEAM MEMBERS

JAY MAJITHIA
VENTHAKRISHNAN ADIKESAVAN

GUIDED BY

PROF. NASSER FARD

Introduction

We focus on performing multiple linear regression for this dataset. We use multilinear regression to find the linearity of this dataset. Multilinear regression is used to determine whether the dataset analyzed from this algorithm is a linear or non-linear model (i.e. Polynomial, Exponential). In this project, we use a package such as pandas, matplotlib, and scikit-learn are used to read data, visualize data, and train models.

Dataset description

The dataset that we are going to use is from Kaggle. It has 13 variables (Pickup_dt, borough, pickups, speed(spd), visibility(vsb), temp, dew point(dewp), sea level pressure(slp), 1-hour liquid precipitation(pcp01), 6-hour liquid precipitation(pcp06), 24-hour liquid precipitation(pcp24), snow depth(sd), holiday(hday). The dataset provides Jan 2015 to June 2015 six months of data which has a total of 29,101 observations.

	pickup_dt	borough	pickups	spd	vsb	temp	dewp	slp	pcp01	pcp06	pcp24	sd	hday
0	2015-01-01 01:00:00	Bronx	152	5.0	10.0	30.0	7.0	1023.5	0.0	0.0	0.0	0.0	Y
1	2015-01-01 01:00:00	Brooklyn	1519	5.0	10.0	30.0	7.0	1023.5	0.0	0.0	0.0	0.0	Y
2	2015-01-01 01:00:00	EWR	0	5.0	10.0	30.0	7.0	1023.5	0.0	0.0	0.0	0.0	Y
3	2015-01-01 01:00:00	Manhattan	5258	5.0	10.0	30.0	7.0	1023.5	0.0	0.0	0.0	0.0	Y
4	2015-01-01 01:00:00	Queens	405	5.0	10.0	30.0	7.0	1023.5	0.0	0.0	0.0	0.0	Y

Fig 1: The partial view of the dataset

Multiple linear regression

There are 12 variables to the problem to predict the number of pickups that occurs on a specific time period of a day. We had 3 categorical variables (pickup_dt, borough, and hday). The pickup_dt variable is in date+time format so it needs to be transformed in the form a date and time. After that, we found there were 3043 missing values in the borough so while converting the data in the numerical value by considering the missing value as a location of its

own. The last categorical column is hday which is transformed in the numerical format (i.e. yes-1 and no-0).

The first thing we did with the clean dataset is to check the data correlation and linearity between variables. In this dataset that we find out from the output result that the relationship between variable is nonlinear since the data was not fitted properly using the Linear regression Algorithm. So, after our detailed research we found that Random Forest Regression algorithm and it fitted the regression model with a coefficient of determination (R^2 score) 91%.

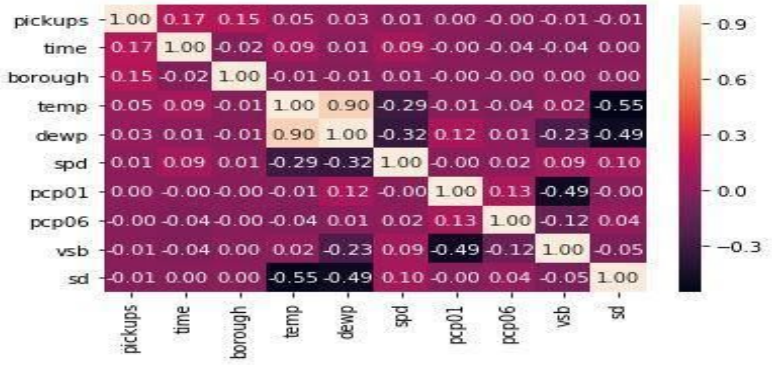

Fig 2: Top 10 collinear factors with the Pickup

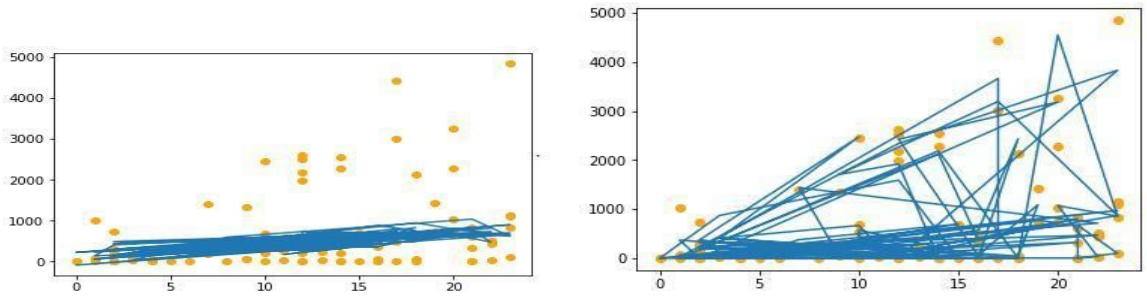

Fig 3: Shows the linear fitted model(left) and non-linear fitted model(right) blue line is the predicted line and the orange points are the actual value.

The question we had which factor affects the most in predicting the pickups on a specific day and time. The correlation heatmap suggested that Borough and time as the most significant factor which raised the question for the factor of the significance of the presence of holiday. On further analysis, it was found out that the data is an unbalanced dataset and it has number of

non-holiday day data than compared to the data for Holiday with a ratio of (27:1). On performing the balancing of data it was found that the borough, time and hday are the most significant factors for making the decision for the number of pickups.

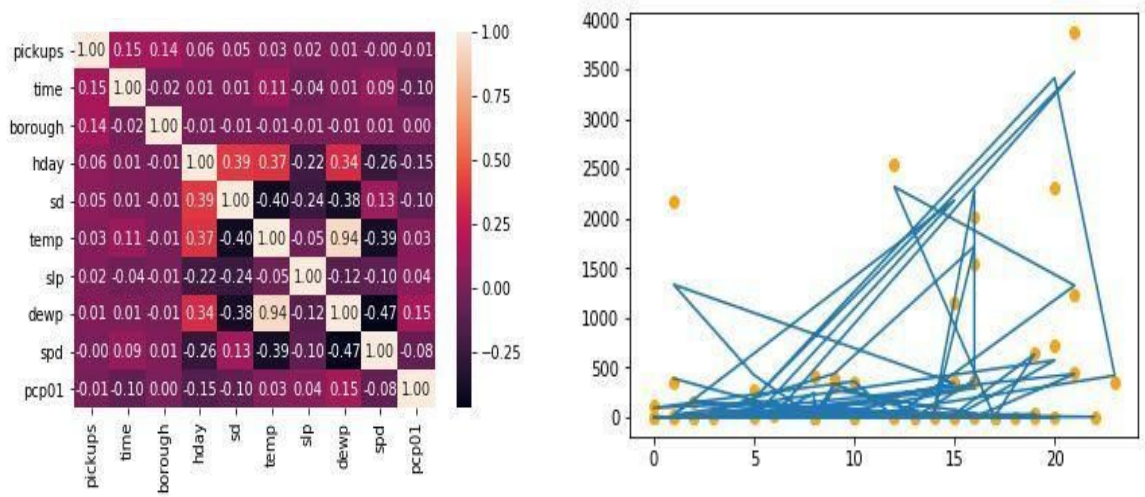

Fig 4: The Left figure shows the correlation of the pickup with its top 10 variables and time is one of them. The right figure shows the fitting model which is improved compared to the unbalanced dataset.

Conclusion

With the given condition we can predict the number of pickups and the data is more than adequate to perform regression analysis. If the dataset would have been equally distributed the significance of some factor and its effectiveness would be identified more significantly. But for the given dataset the test results suggest that with few factors such as Borough, time and holiday are much more significant and can be used to get similar results.

References

Probability & Statistics for Engineers & Scientists 9th edition by Ronald E. Walpole, Roanoke