**ITCS 6100 Big Data Term Project: NYC Parking Tickets**

**EXECUTIVE SUMMARY**

Arpitha Narasinh Alur, Ron Abraham Cherian, Saketh Gourishetty, Shama Zabeen Shaik, Sneha Ramesh Watharkar, Snigdha Bose  
Department of Computer Science  
University of North Carolina CharlotteCharlotte, United States

[aalur1@uncc.edu](mailto:aalur1@uncc.edu), [rcherian@uncc.edu](mailto:rcherian@uncc.edu), [sgourise@uncc.edu](mailto:sgourise@uncc.edu), [sshaik3@uncc.edu, d](mailto:sshaik3@uncc.edu)[swathark@uncc.edu](mailto:swathark@uncc.edu), [sbose1@uncc.edu](mailto:sbose1@uncc.edu)

**Problem Description**

New York, being the city that never sleeps and has way too many visitors from all over the world coming over to look at the places in and around it. With tons of people travelling and the sizes of the parking lots being relatively small to accommodate all the visiting and local vehicles, the ticket issuing system has been enforced in and around the busiest areas of New York. Our project deals with one of the datasets from the ticket issuing systems of the New York and we would be aiming at answering few of the common questions people might have with respect to this scenario.

**Dataset Description and Data pre-processing**

Data is collected by the NYC Department of Finance each year from 07-01 to 06-30. We have made use of the 2016 year’s data from Kaggle. The data initially had 10M records. We have randomly selected 65,000 records and started with pre-processing. In the data preprocessing stage, making use of python script, we have first dropped the columns that had all missing values, then deleted the rows with missing values and dropped the rows that had unrelated values. This preprocessing step has left us with around 27,000 records, which we fed into the Excel Miner for further stages of analysis. We binned the data as per requirements and partitioned it into training, testing and validation datasets which we used for the further stages of analysis and prediction.

We started with exploratory data analysis (EDA) using python, to find out the possible solutions of the hypothesis questions we had. The EDA has left us with results like:

1. Maximum violations on Mondays’ in K county.
2. Passenger plate types are causing maximum violations on Mondays’.
3. Overall maximum violations were observed in the month of June in county K.
4. Violation 21 was observed to be caused the most by the passenger plate vehicles.
5. Maximum violations were observed in county K during mornings.
6. Maximum violations are being caused by passenger plate types during morning.

To understand the insights of the hypotheses better, we have applied certain machine learning algorithms on the pre-processed data using XLMiner.

**Modelling**

We started with*Association Rule Mining***,** but the rules generated from this method, were not well aligned with our hypothesis**.** So, we choose the second method which was hypothesis dependent. *KNN, Random tree, Multilinear Regression* were the methods that were generated separately for the respective hypothesis but they too weren’t giving out the results with good precision and accuracy. We then choose **Naive Bayes Classification**, which helped us in deriving the solutions for all the hypotheses with better accuracy and precision.

**Conclusions**

Overall we are able to predict that **Kings County** (**Brooklyn**) is the area most prone to violations especially on Friday, Monday and in the months of **September, August. And it was also observed that Passenger plate types are the ones causing most violations and during the day time(AM).** The NYC police department should do something so that at these times at these places violations can be reduced.

**Recommendations:**

We recommend that the policy makes should change the trivial violations and also recommend police force to patrol more in places like K county during mornings on Fridays and Mondays to help the passengers. We also recommend installing blackspots on vehicles.