ANALYSIS OF CHARLOTTESVILLE PARKING TICKETS

A PREPRINT

Jacob S. Rodal
jr6ff

Harun Feraidon
hf6mw

Kaan Katircioglu
kk2dh

March 19, 2019

ABSTRACT

This project aimed to find a strategy for where citizens of Charlottesville should and shouldn't park based on local parking ticket data as well as model how the distribution of tickets in Charlottesville evolve over time. We first cleaned our data which involved geocoding and date/time format standardization and then we plotted a map that shows the distribution parking tickets in Charlottesville. Classification models were used to predict whether or not a parking ticket is appealed based on various factors leading to the initial ticketing and regression models will be used to model the distribution of tickets in Charlottesville over time. Through this project, we hope to gain valuable insights about parking ticket patterns in Charlottesville.

1 Motivation:

The problem we are tackling involves being strategic with parking around Charlottesville. Because Charlottesville is a college-city, many tourists and visiting families will attend the city for a short duration. During their stay, because they are not completely fluent with parking regulations around the city, people will often struggle in finding parking spots. Some areas will be more heavily regulated, resulting in a greater amount of parking tickets. On the other hand, other areas might be less regulated, resulting in a lower amount of parking tickets. We hope that by exploring this dataset, we will find a strategy for where one should park and should not park. We will do so by exposing how certain features are related to the chances of getting a ticket, such as the area, the time of day, and what violation may be related to the ticket. Using this information, we can also determine what factors contribute to successfully appealing. Citizens of Charlottesville could use this information to decide whether or not they should bother appealing a ticket.

We would also like to explore how the distribution of tickets evolve over time. The data set provides us with all of the tickets distributed throughout the city over the past 20 years. We can create a regression model to gain insight into how ticketing behavior by the police has changed over time. Using this, we can predict what ticketing behavior will be like in the future, which could be used by Charlottesville officials to inform future decisions regarding the construction of parking garages, parking lots, etc.

2 Method

Before creating our models, we first needed to clean our data. One thing we did was geocode roughly 12,000 street addresses into GPS coordinates. To do this, we utilized the google maps geocoding API and stored the results back into our pandas dataframe. After obtaining GPS coordinates, we then modified a feature that contains date information. Essentially, the date data wasn't in a format that pandas works well with and some dates made no sense (e.g., some dates were from the future), so the date feature was transformed. Another feature that needed to be cleaned was a time feature. A consequence of the dataset being updated for 20 years is that the format in which time data is reported has changed multiple times throughout the years. The feature was put into a single format that pandas can understand. It was important to clean the date and time features because it will be helpful later on when we perform regression and classification that depends on these features. Once all of the data was cleaned, we plotted the GPS data over top of a basemap of Charlottesville so we could get a better understanding of where all of the tickets occured in

Charlottesville. We will show different trends on this map as we discover how the distribution of tickets evolve over time in Charlottesville and what factors lead to sucessful ticket appeals. Here is a portion of what the map currently looks like:

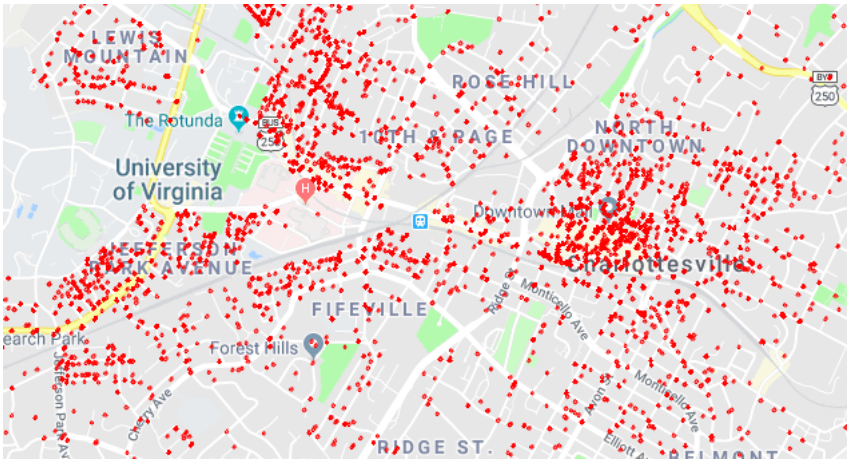

Figure 1: A portion of the map of Charlottesville

When our program was finally ready to be trained, we experimented with several different models to see how they would perform against our data. Firstly, because our problem involved classification techniques, we fit our data among the algorithms we have learned so far: Logistic Regression, Linear SVM, and Decision Tree. We also fit a Random Forest model on our data to get a perspective on how ensemble learning would perform. Finally, because we have heard of these algorithms being used in classification, we also tried using a Neural Network, Naive Bayes, Gradient Boosting Classifier, and Nearest Neighbors to compare their performance against our other classification algorithms.

3 Preliminary Experiments

We intend on studying the methods used by these individuals to inform our research. Furthermore, we will compare our results to their results in an attempt to verify our results where applicable and compare different parking ticket trends in different areas.

Both the labels and an important feature, violation description, were converted to categorical info with one hot encoder.

During the preliminary experiments, the entries with appeal status, labels, "pending" were ignored. Many of the potentially useful features were also ignored, due to additional cleaning and tuning required. Until this checkpoint, the main goal was to come up with at least one working model to make prediction on whether a ticked will be appealed or not.

As explained previously, we tested eight different classifiers and the results are shown below.

Classifier	Train Score	Test Score
Gradient Boosting Classifier	0.661334	0.603116
Random Forest	0.723428	0.593286
Decision Tree	0.723428	0.589206
Linear SVM	0.577119	0.561387
Nearest Neighbors	0.646680	0.560645
Neural Net	0.561352	0.547663
Logistic Regression	0.549898	0.539132
Naive Bayes	0.548414	0.538205

4 Next Steps

A potential problem with one hot encoding was that the violation description had too many different entries, making the one hot encoded matrix too wide. In the following weeks, we'll be rearranging features to a more reasonable number of different, useful features.

For instance, a permit violation and a time violation would capture majority of the classes in violation permits.

We are also going to re-incorporate geocoded coordinates and dates that are more useful. At the moment, we're suspecting that the latitude-longtitude is just interpreted in magnitudes in a similar fashion to GDP-per capita where the difference is binary: larger or lower values. However the locations and potentially times have more subtle relationship with the labels. We might consider categorizing GPS coordinates into a small number of specific locations, such as the corner, the downtown mall, etc. This might be more useful than raw gps coordinates, since GPS coordinates that are very close to one another might be very similar in other ways.

We found that Gradient Boosting Classifier, Random Forest, Decision Trees, and Linear SVM models performed the best among other models. However, because we have not yet adjusted to any hyperparameters, and because Random Forest and Decision Trees model performed the exact same, we will likely narrow our models used to be Gradient Boosting Classifier, Random Forest, and Linear SVM.

Lastly, we also want to model how the distribution of tickets evolve over time, which we have not had time to do yet, as data cleaning and classification took most of our time.

5 Contribution

- Jake Rodal: Performed data cleaning such as geocoding and date/time standardization, handled map creation, helped with classification tasks. Imported many wonderful modules.
- Kaan Katircioglu, Son of M.D.: Presented insights on how to prepare data. Handled feature encoding. Helped Jake with modeling. imported pandas, sci-kit learn, graphicx. Orginzed the meetings.
- Harun Feraidon: Contributed to creating the data pipeline. Delivered pizza 1 day. Helped Kaan help Jake.