IST 5520 Fall 2021

Group 9 Project

Analysis and Prediction of Baltimore Towing Trends

1. **Introduction to the Problem:**

Over the past 12 years in Baltimore, Maryland, there were more than 265,000 instances of towed vehicles. Baltimore kept detailed records including the vehicle make, model, the tow contractor, the location of the vehicle at the time of the tow, the price of the tow, and where the vehicle was stored. Suppose that the City of Baltimore is performing an audit to ensure that their towing practices are unbiased. They are interested in determining if they are unfairly favoring one towing company, if they are unfairly towing specific models and makes of cars, or tow cars from other states more often than others. They also want to predict if a vehicle will be auctioned based on the data provided in the dataset. They hired us, a data science contractor to analyze their data and present the outcome as a neutral party.

## **1.1 Business questions to be answered**

To solve this problem, some of the questions that will be analyzed to check if the city of Baltimore has been biased in towing are:

* Is the city of Baltimore unfairly favoring one towing company over the other competitors?
* Is the city of Baltimore towing a specific type of a vehicle unfairly more than others?
* Is the city of Baltimore towing visiting and travelling visitors disportionately to local residents?
* What is the predicted trend of vehicles being auctioned in the future?

The basis of this problem is the recent news articles and publications being circulated in and around the city of Baltimore stating that the city of Baltimore is being unfair with towing in the city with some sources accusing the city of corruption in towing. Articles and reports related to bias and corruption problems within the city of Baltimore include:

* <https://foxbaltimore.com/news/local/ig-report-city-towing-contractor-sold-towed-vehicles>
* <https://foxbaltimore.com/news/local/in-depth-maryland-leaders-and-corruption>
* <https://www.baltimoresun.com/latest/bs-md-ci-officer-guilty-20111208-story.html>
* <https://www.baltimoresun.com/maryland/bs-md-towing-history-20110223-story.html>

The dataset used in this project has met all of the criteria required for the project. It has 39 variables and 265,816 observations so it is large enough to analyze using data science techniques. The dataset is not clean as it contains many missing values and the time data will need to be reformatted for ease of analysis along with many other adjustments. As far as we have been able to discover, the dataset has not been analyzed by anyone on the internet. The dataset appears to be untouched. This could possibly be due to its restricted access.

**1.2 Variables**

In this project, the following variables will be used in the analysis to answer the business questions: TowedDateTime (Date, The date and time the car was towed.), PickupType (String, The reason the car was towed.), VehicleType (String, The type of the car.), VehicleYear (String, The year of the car.), VehicleMake (String, The manufacturer of the car.), VehicleModel (String, The model of the car.), VehicleColor (String, The color of the car.), TagState (String, The state the car was registered in.), TowCompany (String, The company contracted to tow the car.), TowCharge (Double, The amount charged by the tower to tow the car.), TowedFromLocation (String, Where the car was when it was towed.), ReceivingDateTime (Date, The date and time the car was received at the yard.), StorageYard (String, The storage yard the car was towed to.), StorageLocation (String, The location of the storage yard.), HoldDateTime (Date, The date and time the car began to be held.), HoldReleasedDateTime (Date, The date and time the car was released.), HoldReleasedNotifyDate (Date, The date and time they notify the owner.), RemovedFromYardDate (Date, the date and time the car was removed from the yard.), Status (String, The status of the car (released, auctioned, etc.), ReleaseDateTime (Date, The date and time the car was released.), ReleaseType (String, How the car was picked up/realeased.), TotalPaid (Double, The amount of money paid.)

1. **Sourcing and Collecting the Data**

The city government of Baltimore openly provides much of its data collection for access free of charge. They host a large csv file on ArcGIS which includes descriptions of the variables included and an open license for use of the data. We did find that we had to have an account with ArcGIS/Baltimore to download the data, but it was easy enough to copy and paste the data into a new excel spreadsheet, as it was already in csv format. The dataset has almost 265,000 entries stretching over a period of more than 11 years and still continuously being updated with new records. The data in full at the time of our processing is currently in the referenced document ‘Towing.csv’, but can also be found at:

<https://data.baltimorecity.gov/datasets/baltimore::towing/explore>

<https://data.baltimorecity.gov/datasets/baltimore::towing/about>

1. **Data Manipulation and Refining:**

The dataset in its raw form contains many missing values, strings that are misspelled, dates and times of various formats, and many categories that will need to be modified.

We immediately dropped 18 columns from the total of 39 as they contained many missing values, or we deemed them wholly unimportant in the analysis due to their specificity. We changed several of the time columns for easier analysis, including changing year to a numeric variable, and using python’s datetime package to better define the time periods in different time columns.

We then moved on to detecting and removing null or NaN values and outliers. We first determined the total number of records that included null values, and determined that 253,000 of the 265,000 entries had at least 1 null each. To hopefully negate this we dropped 3 columns that contained a large number of null values and that we deemed were unimportant for analysis - they were HoldDateTime, HoldReleasedDateTime, and HoldReleasedNotifyDate, which were rather repetitive, as what rare entries there were often occurred within hours of the cars being impounded or released, which were included in other variables. We also added a category called ‘unknown’ to represent entries in records where the data was unknown.

After this manipulation we were able to bring down the total number of null entries to 135,000, meaning we would be able to retain approximately half of the dataset after dropping null values. To detect outliers we used boxplots on the numerical variables such as ‘TowCharge’, and ‘TotalPaid’ which had some very large outliers that could have skewed analysis.

1. **Data Summarization and Visualization**

Summarization of the data took up much of the preliminary analysis because the dataset includes variables that have many categories - some of them misspelled or repeated. For instance the ‘TowCompany’ variable has 23 categories despite 8 of those categories being repeated or variations of names already appearing as categories.

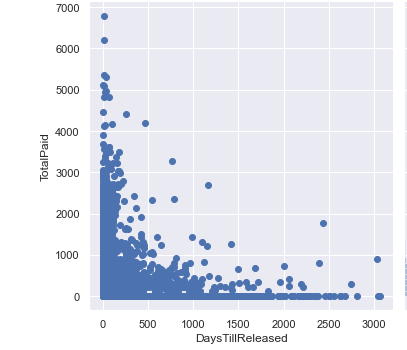
A very interesting result was the count of how often a towing service was used, which provided insight into the city’s choice of contractors. Just 3 contractors accounted for nearly 50% of the total usage by the city. Another interesting result was the contingency table describing how each company charged use of their services, which showed that all instances had a rate of $130 charged no matter the contractor.

For further insight into the data, we considered the different reasons for a pickup, and attempted to determine if the towing companies were selective in their choices for pick ups. This was done using cross tabulation between ‘PickupType’ and ‘TowCompany’.

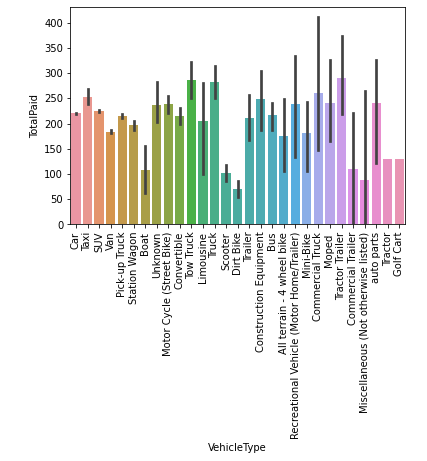
Pickup types ranged from accidents, abandonment, illegal parking, police action, narcotics, recovering stolen vehicles, street cleaning, and scofflaws.

The results showed that some companies did pickup for more of one type than others, such as Aaron’s Automotive Services and Frankford Towing being chosen more often for accidents and other companies such as Mel Del’s Towing and Recovery and Universal Towing being used for recovering stolen vehicles. Interestingly, the city’s towing branch were the ones who took on scofflaw’s and abandonment - this may reflect the city’s intent to use abandoned vehicles for their own intent, including possibly auction or repurposing - this could include potential corruption - an important part of analysis could include determining if the city used its resources to sell of abandoned vehicles or vehicles owned by scofflaws for profit.

We also made a contingency table that described the total number of vehicles towed by each year that the car was made and found that the distribution was roughly normal, with the average year being around 2000. Next we visualized what variables we could. This included making a violin plot of the distribution of tow charges to see how frequently each amount was charged, a bar chart of ‘ReleaseType’ which described why the vehicle was released from the city’s impound lot, a barplot of time impounded vs the total charge paid, a barplot of the type of vehicles impounded, and also a barplot of the vehicle type versus the total amount paid for the tow.



The time impounded is not evidently significant towards the total price spent. Most high cost payments occurred over a short time period, as opposed to a longer period. However, this could reflect that the city charges higher prices for people to recover their vehicles quickly.



The type of vehicle impounded seems to have a significant effect on the total charge. Although it is hard to tell why this occurred, it could have to do with why the owners of certain types of cars would let them stay impounded longer, or if they initially cost more to release depending on type.

1. **Regression Analysis**

Prior to the regression analysis, the dataset was already cleaned up in previous steps. An important step at this point was to create dummy variables in place of the categorical variables. To further process the data for the regression analysis, we prepare for independent variable (X) and the dependent variable (y) using the existing variables after initial data manipulation and refining. The dependent variable in this analysis was “auctioned” release type and the independent variable was made to be a concatenation of all the remaining variables, dummy variables and a constant but excluding “auctioned” release type. For the analysis, we described the linear model using ordinary least square regression (OLS), fit a full linear model with the data and summarized the results. This generated the OLS results including the Coefficient of determination (R2), intercept, coefficient of the multiple independent variables, and many others. These results helped us predict the release type of vehicles towed.

1. **Predictive Analysis:**

Our predictive analysis is going to focus on attempting to determine future trends of Baltimore’s towing practices, particularly on how it auctions off different types of vehicles depending on their characteristics, and how much the city aims to profit off of its future auctions. Other possible predictive practices could include determining how long different types of vehicles are impounded, and seeing if there is any causation between the type of vehicle and its initial impound discharge cost and its increase over the time it spends impounded.

Prediction can be done using regression models - however our models are heavily categorical and so will likely do better with algorithms more suited for categorical data, such as kNN, Naive Bayes, decision trees, and neural networks.

The easiest to implement is kNN, which is suited very well for predicting categorized variables, as it is nonparametric and a black box, so it doesn’t require any understanding of how the variables mechanistically interact. Using the sklearn neighbors module allows for quick use of a kNN classifier that can be used to predict new data.

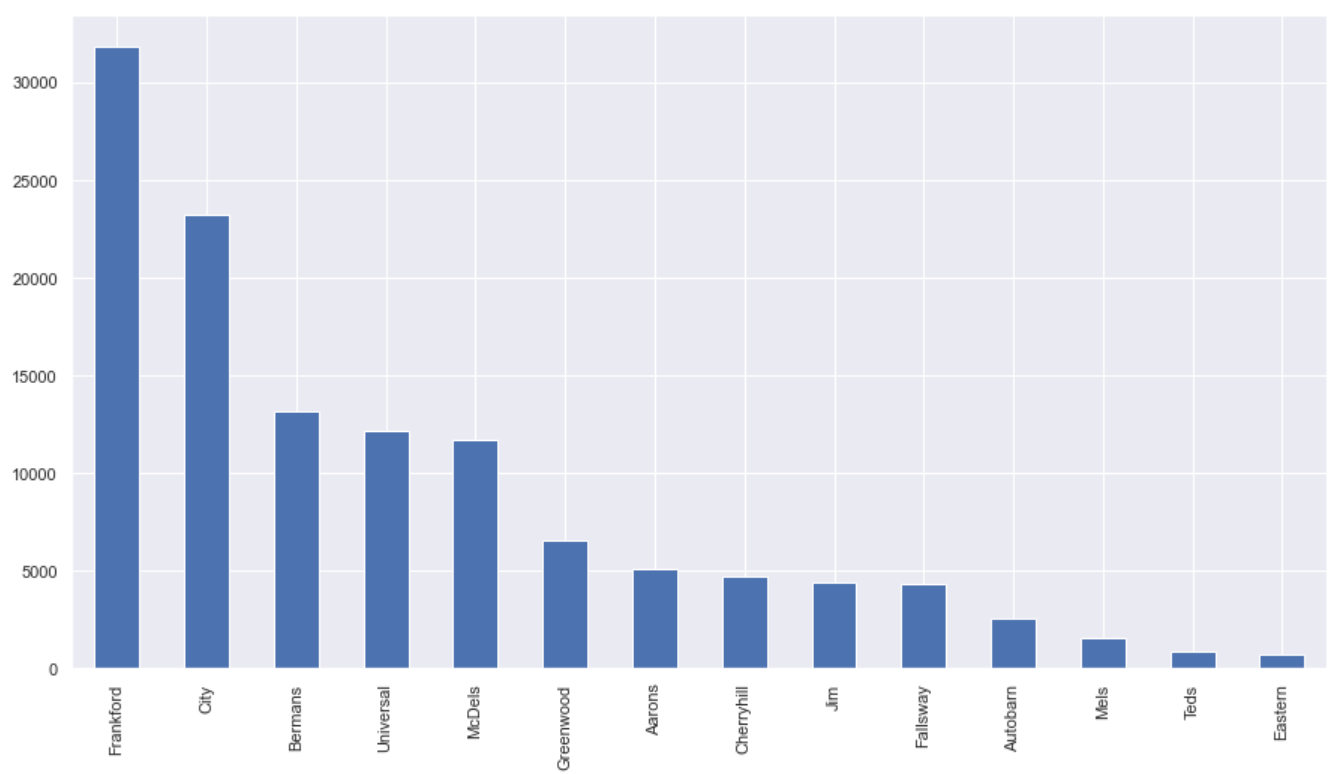
Next we can test a Naive Bayes model, which simply relies on the propensity of a record to be categorized into a specific class to predict which class a record will belong to. However, this does require that we have records that are similar enough over all predictors to provide a basis for future predictions - the main problem here is that we must have a large number of records, which we do. We also implemented decision trees and random forests, as well as neural networks.

Our predictive analysis focuses on classifying records into the class of ‘Auctioned’ as either yes or no, so we will have a binary outcome.

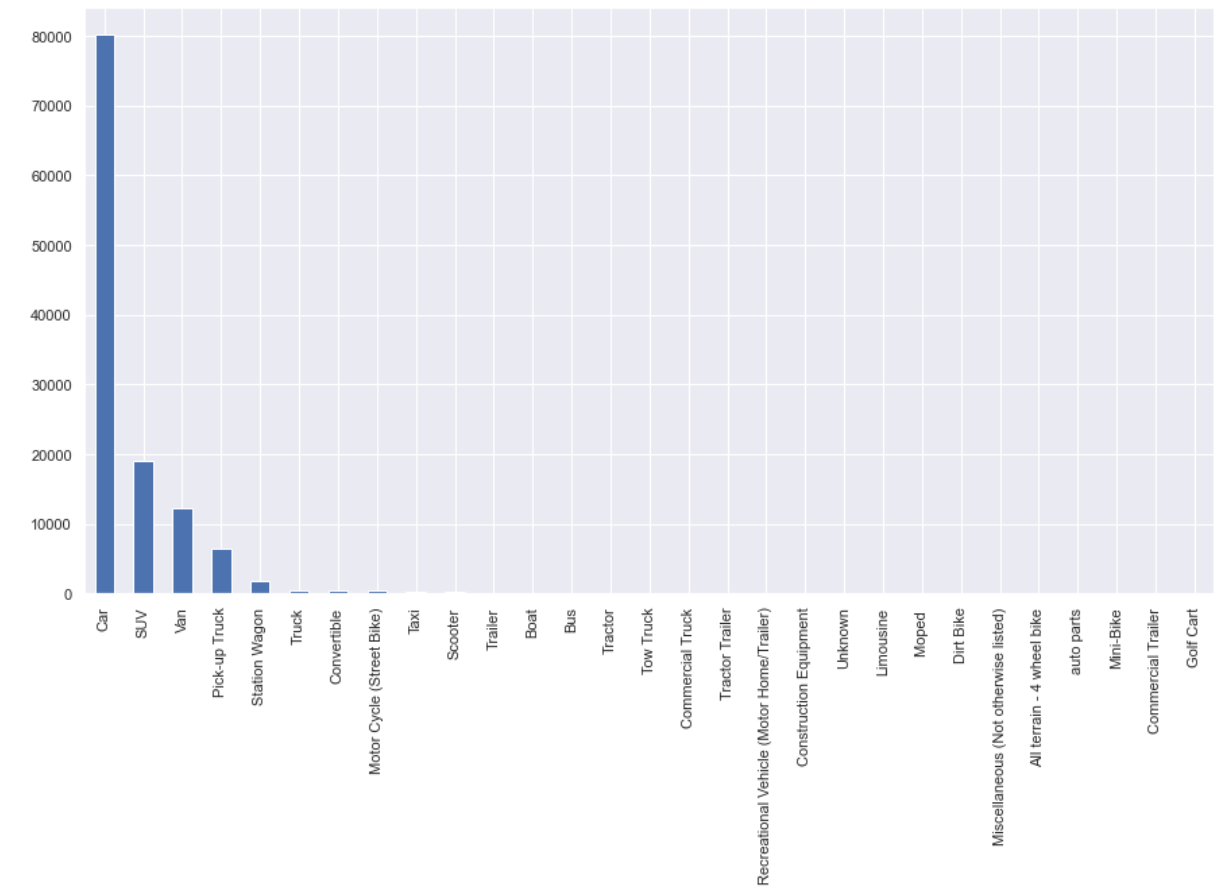
However if we also wish to calculate the city’s future profit off of auctions we can simply use a multiple linear regression model.

1. **Summarization of Results**

The first problem in this project was to determine if the city of Baltimore was unfairly favoring one towing company over the other competitors. When looking to solve this problem, we saw that Frankford Towing LLC had the most towing instances by a significant number. This was followed by the City and Bermans with the second and third highest towing instances respectively. We saw that this company, which had 32,812 towing instances had over 8000 more towing instances than the City which also had over 9000 more instances than Bermans. The remaining companies had from about 1000 to about 13000 towing instances. At this point we can not say with certainty that Frankford Towing LLC is being unfairly favored. However, this could be possible due to their significantly high number of towing of instances as compared to other towing companies. If not, then possible reasons could be because they might have a contract with the city or because they are a bigger company with multiple locations and much more towing trucks. Since we do not have the details to look at those details, we advise this should be further looked into.



We needed to check to see if the city of Baltimore was towing specific types of vehicles more than others. We found that cars were being towed the most way more by far than any other type of vehicle. Since cars are the most popular type of car being driven, that may seem like the obvious choice. So to further dig into this, we wanted to analyse to see if other types of vehicles were being unfairly treated even though they may not be as popular as cars. We could see that smaller vehicles like dirt bikes and scooters have the lowest tow charge and some of the biggest vehicles such as tractors and trucks have the highest tow charge. This is because more time, labor and resources are spent when towing larger vehicles than smaller vehicles. Therefore we determine that there is no unfair treatment to vehicle types. If vehicles were being treated unfairly, smaller vehicles would have had higher charges and vice versa.



The next problem was to determine if the city was targeting visitors disproportionately. From our results we realized that had the vast majority of occurrences with vehicles from neighboring states being towed much less frequently. Based on this result, it was obvious to us that the city is not unfairly targeting visitors to have their vehicles towed.

In the predictive analysis, KNN, Gaussian Naive Bayes, Bernoulli Naive Bayes, Decision tree, random forest and artificial neural network predictive models were used. A 20-80% split was performed on the dataset for training and testing the models. Accuracy, AUC and Cohen kappa metrics were used to evaluate the performance of all the models for comparison. We found out that Artificial NN has the highest accuracy and AUC. To predict if a car is auctioned, we recommend using ANN since it has an accuracy of 99.49%, AUC of 99.51% and recall of 99.55%, the best performance among all the models followed by random forest. In the analysis, we realized that the total amount of money paid and the number of days a car is impounded had the greatest impact on if a vehicle is auctioned.

Our final conclusion regarding the potential corruption within the city of Baltimore is that the city does not exhibit any extraneous amounts of corruption. However, further analysis into their choice of towing contractors is needed. If we had access to their decision making process we could improve our analysis and decision.