Final Project Paper

Vincent Morin

12/9/2019

Contents

Introduction

Aim: Conduct analysis on a dataset utilizing the machine learning tools and methods covered in the class material, and articulate the methods, processes, and findings in report-format.

The Problem: Conduct analysis into the relationships between average tip-size per day/hour and other external variables such as weather, time/day, number of passengers, etc..

Goals:

- Use analytical machine learning tools and methods to predict the tip-per-ride of New York City taxis.
- Run scenarios with different variables, and data aggregation to compare and understand results.
- Speak intelligently to the different methods used, the logic behind the different scenarios, and what lessons were derived from this project.
- Uncover our best predictor variable.

Report Roadmap: This report will cover work conducted on two scenarios, both derived from the same source data, seperated by scaling methods. The first scenario aggregated taxi data based on day, and includes daily recorded weather variables. The second scenario follows similar logic, but aggregates around hour-intervals. The report is subdivided into sections, reflective of the work conducted on the data. Sections include: introduction, background, data sourcing, data wrangling, data analysis, results, and conclusion.

Background

New York City is home to one of the most faceted and complex urban infrastructure systems in the world. In 2013, it was estimated that daily commuters into the city nearly doubled the total population on the island, from 1.6 million to 3.1 million (Metcalfe, 2016). This massive transit system provides a considerable amount of data; extensive analysis has been conducted to better understand how it operates, and mitigate it's friction. In 2011 the Green Taxi was created following analysis by the city which showed a lack of taxi coverage in Upper Manhattan and the outer boroughs. This project will focus on data collected on all taxi rides

conducted by New York's Green Taxis in 2016.

Data Sourcing

1. 2016 Green Taxi Trip Data

This dataset was downloaded from the **New York City OpenData** site; an online, publically available source for datasets on infrastructure and public systems operated by the City of New York. The 2016 Green Taxi Trip Data contains recorded information from all taxi trips completed by New York City's Green Taxi service. The data was sourced and collected on behalf of the NYC Taxi and Limousine Commission (TLC), and was uploaded to the NYC OpenData website on September 10, 2018 (NYC OpenData, 2018).

2. Weather data in New York City - 2016

The 2016 weather dataset contains daily information on weather out of the Central Park weather station. It was originally sourced from the National Weather Service Forecast Office's online data portal (accesible **here**), and was downloaded from **Kaggle**. Variables in the dataset include: date, max temperature (in Fahrenheit), min temperature, average temperature, preciptation, snow fall, and snow depth.

3. NYC Hourly Weather Data

Like the previous dataset, the NYC Hourly Weather Data dataset contains time-based recorded weather data out of New York City. Unlike the previous dataset, this data was recorded hourly. The dataset was accessed via **Kaggle**, and was originally sourced from Wunderground, a for-profit internet weather service.

Data Wrangling

Much of the different methods and tools in statistical research and machine learning revolve around expanding, understanding, and experimenting with different independent variables, ultimately testing what fits well, and what doesn't fit so well in a model. Data wrangling for this project was conducted in three parts. The first part covered intital wrangling and preparing of the main dataset: the 2016 Green Taxi info. The intention of this seperation was to cut down on processing time and power. The second and third parts include data wrangling in the form of joining datasets, and require scaling relative to the unit of analysis.

2016 Green Taxi Trip Data

In any machine learning project, attention needs to be applied to the size, consistency, structure, and contents of the underlying dataset. Due to the size of the Green Taxi dataset (2.17GB), consideration on what information is useful, and what information can be removed was be taken. In this case, some columns were removed due to: uncertainty/unknown descriptor, constant values, and location. Typically, location could provide very useful information and serve as key predictor variables to the output; think of economic status of neighborhood, or zoning (business, industrial, residential). However, keeping in mind the general purpose and size of this project, location of all rides will be standardized, and assumed to be the New York City metropolitan area, in general.
Columns (variables of interest) retained from the data include:

- lpep_pickup_datetime – A timestamp of when the trip began.
- Lpep_dropoff_datetime – A timestamp of when the trip ended.
- Passenger_count – Total number of passengers during trip.
- Trip_distance – Trip distance in miles.
- Fare_amount – Total fare amount in dollars.
- Extra – $0.50 for rush hour charges, $1.00 for overnight charges.
- MTA_tax – $0.50 New York City added tax per trip.
- Tip_amount – Added dollar value in tip to driver (this will help serve as our unit of analysis).
- Tolls_amount – Tolls added to the trip.
- improvement_surcharge – Added charge to some trips.

- Total_amount – Total dollar cost of trip.

While removing roughly 12 extranneous variables from the data considerably reduced it's size, it still contained roughly 1.64 million rows of data. Keeping in mind that the unit of analysis, and minding the purpose and size of the project, trips without tips were also removed from the dataset. The result of this work reduced the total number of rows down to a manageable size: roughly 690,000. In the following scenarios, the data will be further scaled to and manipulated; creating mutual variables to join with.

Scenario 1

Scenario 1 utilizes weather data collected from the New York Central Park Weather Station. This data is comprised of daily observations of weather variables that include: max. and min. temperatures, precipitation, snow fall, and snow depth. After some primary inspection of the dataset, amendments were made to prepare the data for further analysis, such as changing variables types and replacing inconsistent data observations.

In order to join these two datasets around the time-interval piece of the unit of analysis, the Green taxi data was aggregated to daily observations. To achieve this, observations were grouped by date and averaged. This effectively reduced the size of our taxi dataset from 690,000 rows, to 366 rows. (((See Table 1 for a selection of the first several rows of our cleaned dataset.))) The resulting wrangled dataset includes 19 total variables, and 366 observations.

((Print table of averages:))

Scenario 2

Scenario 2 follows a similar logic and structure as Scenario 1, but requires slightly different data manipulation. Instead of aggregating the Taxi data down to the day and calculating the respective trip averages, trips were grouped by hour. The purpose of this change in the time-characteristic of the unit of analysis, is to allow both the Green Taxi dataset and the

NYC Hourly weather dataset to be joined around hourly-recorded observations.

Additional steps completed during data wrangling in Scenario 2 included work around cleaning the data observations in the NYC Hourly weather data. At the onset, the NYC Hourly weather dataset had 30 variables and 10,481 observations. After further review of the dataset, 11 variables were removed. Several were repeat variables, though in metric measurements. For the purpose of this project, imperial measurements were kept. After joining both datasets, the final working dataset is comprised of 29 variables, and 10449 observations.

Data Analysis

This section covers analysis into, and compares data from both scenarios. Each dataset was indexed; randomly split into respective training and testing datasets. The training datasets cover approximately 75% of the data, and the test datasets include the remaining 25%. Standard practices when conducting analysis on data include: checking for missing values, ensuring data types are cohesive, and refraining from working too closely, or even checking, the test data.

For this project analysis, some additional variables were created and added to the datasets based on the structure, and known variables. For example, for daily measure of tips per ride (Scenario 1), a categorical variable was created to denote the day of the week for each observation. This allows comparisons between the variables: observing if there is any relationship between a given day of the week, and the amount that was tipped (**See Figure 1**). Additionally, similar information was compared against the particular month of the year (**See Figure 2**). Scenario 2 also provides information on what total average tips by hour looks like (**See Figure 3**).

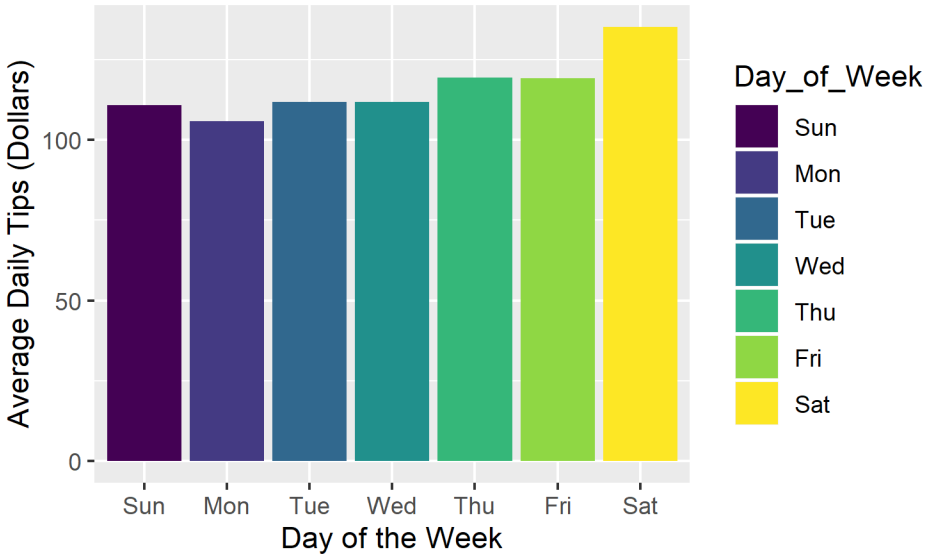

Figure 1: **Average Daily Tips by Week Day**

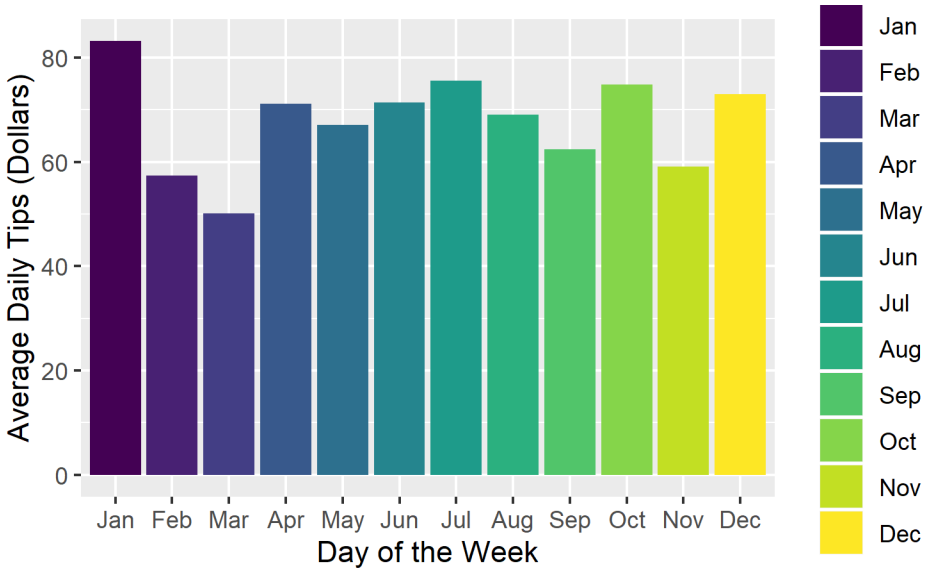

Figure 2: **Daily Tips by Month**

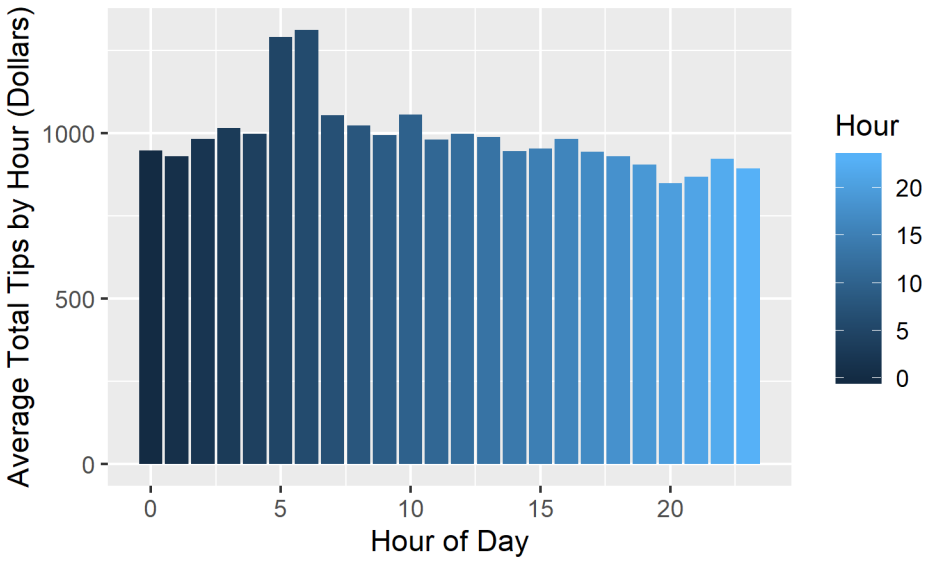

Figure 3: **Total Average Tips by Hour**

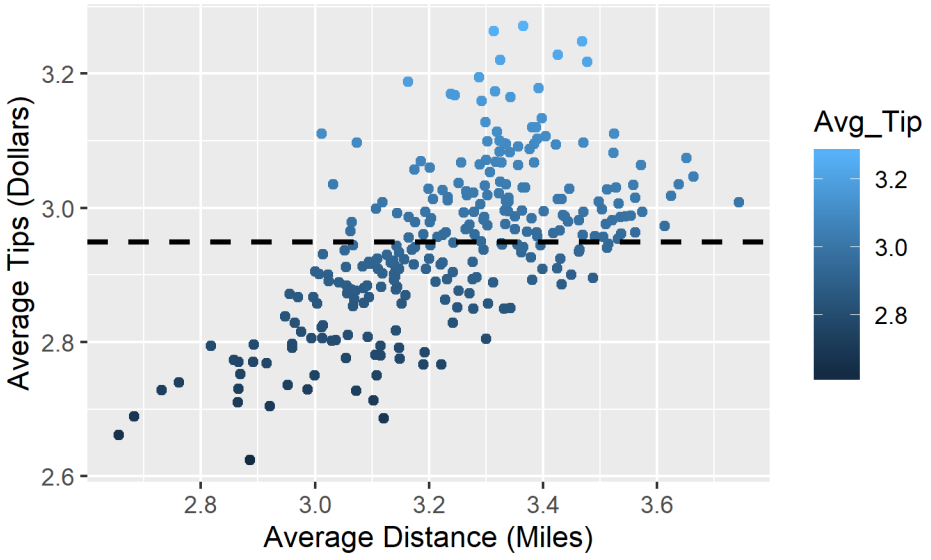

Figure 4: **Average Tips and Distance**

Machine Learning

For this project, data for both scenarios were run through several popular machine learning models, including: linear regression, K-nearest neighbors, classification and regression trees (CART), and random forest. The ultimate goal of these models was to get a sense of how accurately tip-per-ride could be predicted. Questions associated with this practice include: how accurate was each model; which was the most accurate predictor model, which variable was the best predictor, what improvements could be made for better accuracy?

The steps for conducting these processes starts by utilizing the recepies package to prep the data. This includes normalizing the distributions of all variables and observations, inputing any missing values (there were none in this step), and converting dummy (catagorical) variables. Following this, the data are seperated info folds: even portions of the data which assist in training models. In this case, data for both scenarios was subdivided into 5 roughly equal folds.

Results

Below are Rsquared results for the models run on each scenarios. Rsquared is a measure how closely the data fit to a regression line; the closer the value is to 1, the better the fit. In this case, the best model for predicting scenario 1 was the random forest model.

Scenario 1:

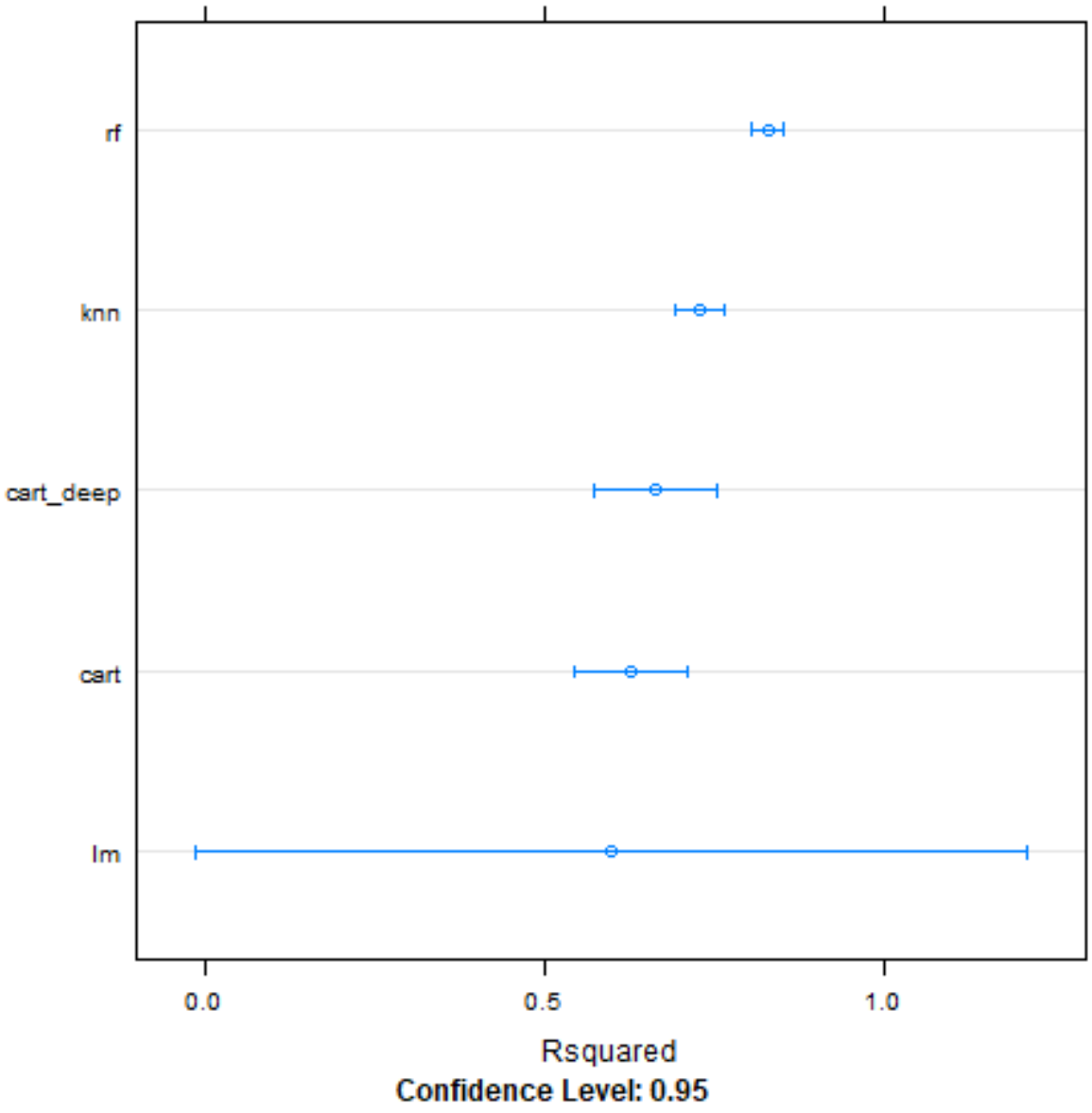

Figure 5: **Scenario 1 Model Results (Rsquared)**

Scenario 2

{I was unable to run the models for scenario 2. While they had previously worked, I recieved an error which stated that there were missing values in my Avg_Tips variable. However, there are no missing observations, and I did not have sufficient time to unpack this issue. }

Conclusion

Ultimately the results of the models ran on both Scenario 1 and 2 were relatively inconclusive. The Rsquared results, as well as the MSE were such that the fits of the models were not accurate predictors of what the expected tip may be. However, the purpose of this project was not to necessarily solve the problem, but instead to apply some of the methods and tools reviewed in this class. In review, running two scenarios was taxing on resources of time and organization. Future projects of this size and composition will need to have better consideration to overall structure and time allocation.

Lessons

- Spending more time on what variables are ultimately useful for conducting analysis, as well as what additional variables could be derived, is essential to a successful machine learning experiment.
- (Learning by doing) More exposure to the material, more practice with the methods provides a better sense of time; answer the question, "how long will this process take me?"
- Ultimately there are aspects of this project that were shortsided due to resource allocation. With more time, additional datasets could have been brought in. There were also opportunities for more exploration around location and spatial data.

Works Cited

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Silge J, Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS*, *1*(3). doi: 10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

Jeffrey B. Arnold (2019). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 4.2.0. https://CRAN.R-project.org/package=ggthemes

Emil Hvitfeldt (2019). textdata: Download and Load Various Text Datasets. R package version 0.3.0. https://CRAN.R-project.org/package=textdata

Grün B, Hornik K (2011). "topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software*, *40*(13), 1-30. doi: 10.18637/jss.v040.i13 (URL: https://doi.org/10.18637/jss.v040.i13).

Max Kuhn and Hadley Wickham (2019). recipes: Preprocessing Tools to Create Design Matrices. R package version 0.1.7. https://CRAN.R-project.org/package=recipes

Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). caret: Classification and Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret

John MetCalfe. "The Many Ways People Commute to New York". September 26, 2016. CityLab. Accessed 12/02/2019. https://www.citylab.com/transportation/2016/09/

manhattan-commutes-port-authority-bus-terminal-capacity-study/501515/