Project 9: Predicting Food Delivery Times

Submitted By: Andrew Stambaugh, Undergraduate Student

Dept. of Mechanical Engineering

Northwestern University

Date Submitted: 12/09/2022

Table of contents

Predicting Food Delivery Times

Andrew Stambaugh

Abstract: This project details the use of the six principles of Mechanistic Data Science to translate a Kaggle dataset of food deliveries to a Random Forest Regression model and Neural Network which can accurately predict the arrival time of food deliveries upon being sent out. Through deep data mining, creative features were engineered to finely optimize the model and accurately represent nonlinear and non-numeric relationships within the data.

Keywords: *Data Science, Machine Learning, Deep Learning, Neural Network, and Feature Engineering*

Introduction

Food delivery is a service which has become increasingly popular in the food industry over the past few years, especially with the COVID-19 pandemic limiting physical contact across the world. As more companies funnel into this market, there is an increased need for fast, reliable delivery times as these best communicate to customers the efficiency of their service. Many companies, including DoorDash, Uber, and more have begun implementing forms of machine learning to provide delivery time estimates. However, can a machine learning model give reliably accurate delivery time estimates when given enough data? This project seeks to answer this question through following the six modules of Mechanistic Data Science (MDS): Multimodal Data Generation and Collection, Extraction of Mechanistic Features, KnowledgeDriven Dimension Reduction, Reduced Order Surrogate Models, Deep Learning for Regression and Classification, and, finally, System and Design [1]. These modules form the crux of this project's methodology, and their applications span across a diverse array of projects, including heart failure classification, audio translation, or even optimizing a sports team model. The main dataset for this project was retrieved from Kaggle and contains over 45,000 entries of food deliveries across multiple platforms and companies. All coding was done in Python through standard libraries such as pandas, numpy, sklearn, and tensorflow. Through completing this project, I sought to not only learn more about MDS as a whole, but also to learn new technical skills in data science such as using tensorflow and coding my first neural network model.

Multimodal Data Generation and Collection

(https://www.kaggle.com/datasets/gauravmalik26/food-delivery-dataset)

This project's dataset was retrieved from Kaggle through the link above and contains over 45,000 entries of food deliveries across different platforms, food companies, and third-party service providers. It contains 19 input features and one dependent variable, the time it takes for the food to be delivered upon leaving the restaurant. The inputs features include Delivery ID, Delivery Person ID, Delivery Person Age (Years), Delivery Person Average Rating (1-5),
Restaurant Latitude, Restaurant Longitude, Delivery Location Latitude, Delivery Location Longitude, Order Date (M/D/Y), Time Ordered Placed (hh:mm:ss), Time Picked Up (hh:mm:ss), Weather Conditions, Road Traffic Density, Type of Vehicle, Vehicle Condition, Type of Order

(Snack, Meal or Buffet), Multiple Deliveries? (0-3), Delivering to a Festival? (Binary), and Urban Density. Table 1 is an example of some of the dataset's features. Upon importing the data, I converted it into a panda data frame in Python.

Table 1: Example Portion of Main Dataset (Raw)

Road_traffic_density	Type_of_vehicle	Weatherconditions	Festival	Urban Density	Time_taken(min)
High	motorcycle	conditions Sunny	No	Urban	(min) 24
Jam	scooter	conditions Stormy	No	Metropolitian	(min) 33
Low	motorcycle	conditions Sandstorms	No	Urban	(min) 26

Mechanistic Features Extraction Part 1

The raw dataset from Kaggle contained improper formatting for many fields. Thus, the data needed to be heavily preprocessed and the proper features be extracted where necessary.

Preprocessing

The dataset required an immense amount of preprocessing before serious feature extraction could ensue. The data cleaning process involved removing NULL values, dividing categorical variables into multiple binary variables, converting all scales to numeric values, universalizing date time formats, and removing any unnecessary strings from numeric entries. Table 2 shows Table 1 after data cleaning.

Table 2: Example Portion of Main Dataset (Clean)

Road Traffic Density	Motorcycle	Scooter	Sunny	Stormy	Sandstorms	Festival	Urban Density	Time Taken (min)
2	1	0	1	0	0	0	2	24
3	0	1	0	1	0	0	3	33
0	1	0	0	0	1	0	2	26

Feature Extraction

After preprocessing, the correct features needed to be extracted from the restaurant coordinates, delivery coordinates, order date, and order time. The coordinates themselves possess little interpretation in a machine learning model besides the distance between them. Using the distance method from the geopy Python library, the geodesic between the coordinates was calculated and the distance recorded. As for the date and time, I used the date time Python library to extract various temporal features from them. These include the day of the week, month, season, hour, rush hour?, night time?, and weekend? These are all features I hypothesized may affect an order's delivery time.

Knowledge Driven Dimension Reduction

Processing the data's features into usable formats consequently increased its overall dimensionality to over 35 features. Due to this, I sought methods of dimension reduction.

K-means Clustering

The first of these methods was to cluster the numerical data with the K-means algorithm in hopes of discovering a latent variable. In Figure 1, the three numerical variables are graphed in a 3D plot, with the colors representing different clusters. As we can see in the plot, the clusters are around the same size but not well separated, implying that they provide little insightful information. This is an issue with the K-means algorithm as it is biased toward equal sized clusters with similar densities. It is visually clear that the data is split where a delivery person's ratings are ~4 standard deviations below the mean. However, this split would not provide clusters which minimize the objective cost function of K-means. Therefore, I sought different methods of dimension reduction and addressed this split later on in Feature Extraction Part 2.

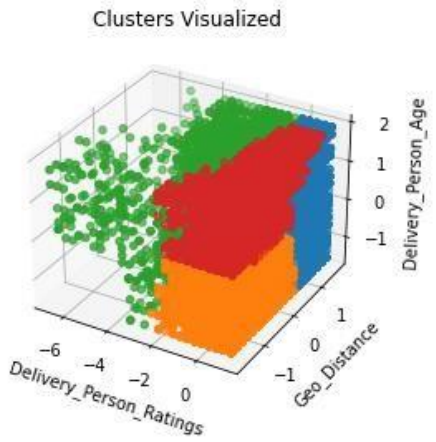

Figure 1: K-means Clusters Visualized

Correlation Matrix

A simple yet effective method for determining a variable's redundancy and relevance is the correlation matrix. In regard to redundancy, I implemented a maximum of 0.6 input-input correlation, but even with this strict criterion, no variables possessed such a correlation together. On the other hand, in regard to relevance, I implemented a minimum of 0.2 input-output correlation, meaning if a variable's correlation with delivery time was less than 0.2, then it was deemed irrelevant and removed from the dataset. With this criterion, the number of dimensions was reduced from 35+ to only 13, including age, driver ratings, traffic level, vehicle condition, geo distance, urban density, multiple deliveries?, festival?, nighttime?, rush hour?, sunny?, fog?, and motorcycle?

Reduced Order Surrogate Model

In the pursuit of dimension reduction, I attempted to find a latent space within the dataset in order to potentially create a reduced order model. However, with the majority of the dataset being binary or categorical variables, finding a proper latent space spanning all of the data necessitated an algorithm beyond the scope of my data science knowledge. Hence, I focused on reducing the order of only the non-binary features.

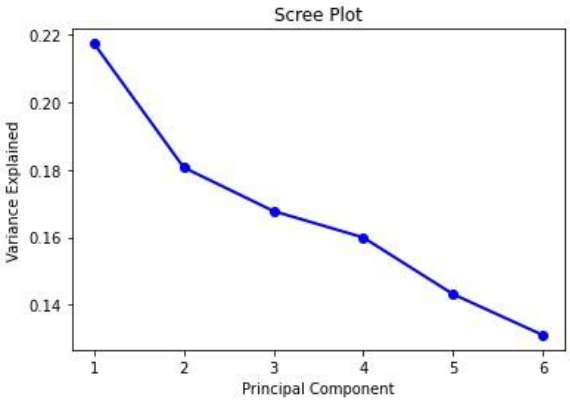

Figure 2: PCA Scree Plot

Principle Component Analysis

Using Principle Component Analysis, the results for a latent space were unsatisfactory. As shown in the scree plot in Figure 2, the first principle component only explains about 22% of the variance in the data. Furthermore, in order to capture at least 95% of data's variance, all 6 components are

needed. This results in no dimension reduction. Thus, a reduced order model was not implemented for the dataset.

Features Extraction Part 2

Although the proper mechanistic features have already been extracted, creative feature engineering must be done in order to best model each feature's relationship with delivery time, specifically non-linear or non-continuous relationships. Deep data mining was done on the dataset, and this section goes over the major takeaways found in this process.

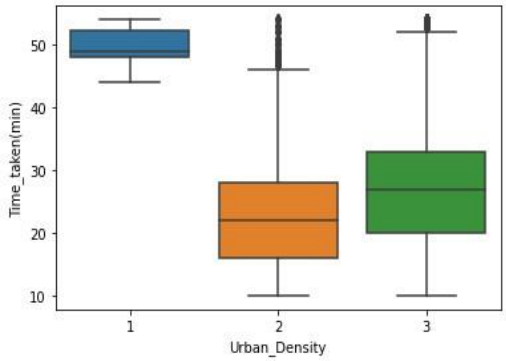

Figure 3: Delivery Time by Urban Density

Lack of Suburban Data

Suburban food deliveries, across all features, behave entirely different than those in urban or metropolitan areas. Not only do they have significantly higher delivery times on average (Figure 3), but features such as geo distance or delivery ratings no longer have any correlation to delivery time. In other words, suburban deliveries need a separate model entirely. With these deliveries only making up 0.3% of the data, I felt it was necessary to narrow the scope of my project to only urban and metropolitan areas as I did not have a sufficient amount of data to create a separate, reliable model for suburban food deliveries.

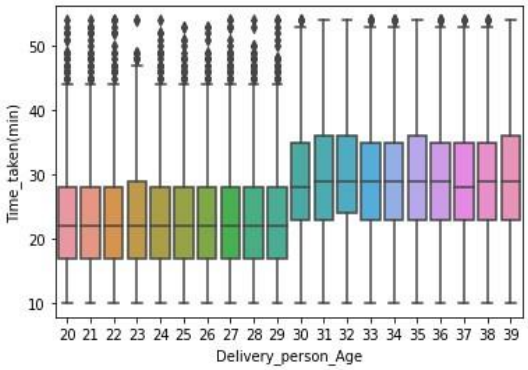

Figure 5: Boxplots of Delivery Time by Age

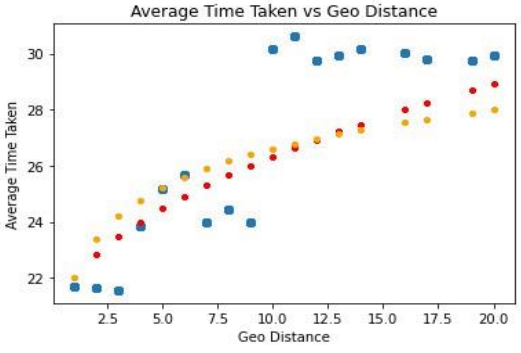

Figure 4: Average Delivery Time by Geo Distance vs Square Root and Log Functions

Dichotomous Variables

Many variables not only showed a nonlinear relationship with delivery time, but also a discontinuous, dichotomous relationship. Examples of this are seen in a delivery person's age (Figure 4), driver ratings, and the geo distance between the restaurant and destination. In Figure 4, the delivery person's age has no effect on the median delivery time until the age of 30, in which it abruptly increases and levels out again. This exact phenomenon happens for driver ratings, as well, at 4.5 stars, except decreasing instead. In order to account for this, these features were transformed into a binary format, splitting the data at the aforementioned dichotomous lines. However, despite having a similar split behavior at values greater than 10 (Figure 5), geo distance still possesses somewhat of a positive correlation on the left side of the split. To account for this, rather than using a simple binary variable, I split geo distance into two variables, geo distance * (geo distance >= 10) and geo distance * (geo distance < 10).

Other Important Relationships

Although not surprising behavior, multiple deliveries possesses a visually linear behavior in relation to the time taken to complete a delivery (Figure 6). This is an important discovery as it supports the hypothesis that under similar conditions, delivery times will have similar values (assuming each delivery in these multiple deliveries is under similar conditions). Simply put, delivery times can be predicted given the correct knowledge and relationships of its conditions. Finally, vehicle conditions have little effect on the delivery time, except when the vehicle is in the worst condition possible. Thus, it was best to represent this feature as a binary rather than a 0-2 scale.

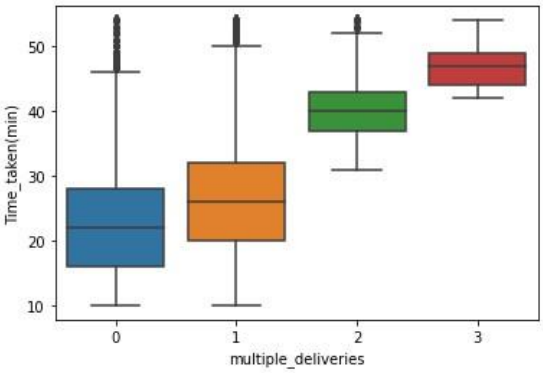

Figure 6: Boxplots of Delivery Time by Additional Number of Delivery on Route

Deep Learning for Regression

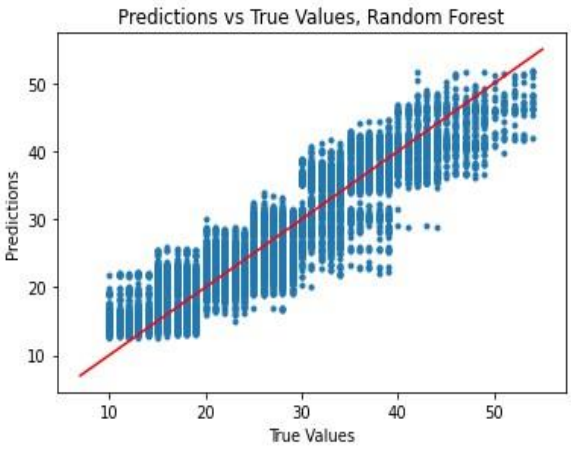

Figure 7: Random Forest Predictions

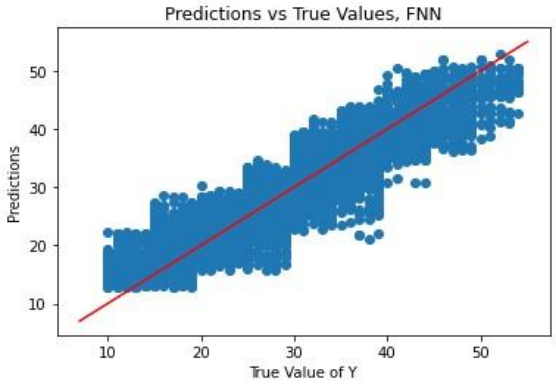

Figure 8: FNN Predictions

Although deep learning more specifically refers to neural networks, I initially implemented a more simple machine learning model with Random Forest Regression. My initial thoughts were that the data in this project was simple, and, thus did not need the more computationally expensive neural network model in order to be sufficiently accurate. Although this proved to be true, I decided to implement a Feedforward Neural Network otherwise for the sake of comparison and to better follow the Deep Learning aspect of this module. *Random Forest vs. FNN*

Other than its simpler implementation than a neural network, I chose to implement this algorithm as it excels in deciphering convoluted, nonlinear relationships as opposed to other forms of regression. Using the RandomForestRegressor() function from sklearn.ensemble in Python, I implemented this model with 500 trees, each with a depth of 11, and only 8 features used, maximizing the R-squared at 0.7433.

For the FNN's implementation, KerasRegressor() was used from tensorflow in Python with one input

layer and two hidden layers, each with 25 neurons and the Rectified Linear Unit activation function. This model's performance peaked around 30 epochs with a batch size of 5 at an R-squared of 0.7278. Though close, the Random Forest model is still slightly more accurate, despite having a simpler algorithm. This is reflected in the plots of their predictions (Figure 7 and 8), as well. While, the Random Forest is limited to integers as that is what it has been trained on, the predictions are generally tighter to the line, indicating that it's more accurate on average. However, as this was my first neural network, there is certainly room for fine tuning and improving that model much further. All in all, both models are respectably accurate and prove that given sufficient features of a food delivery order, it's possible to accurately predict most delivery times.

System and Design

For system and design, the fitted model can be directly implemented into a food delivery service's UI. Upon an order being sent out for delivery, the real time data is recorded by the system and input into this model's algorithm, including the data cleaning and feature extraction process. One concern with this algorithm is that it evidently does not take into account that the conditions of a delivery, such as weather or traffic level, are not constant. However, in such cases, the algorithm can be run again, except instead of the restaurant's location, the driver's current location on the road can be input, imitating a new order being placed. Finally, all data collected by the system can be fed back into the algorithm. In turn, it will continue learning, and its performance will improve over time.

Conclusion

Overall, it is concluded in this project that it is indeed possible to accurately predict the time of a food delivery's arrival through machine or deep learning methods. Whether it be through Random Forest, a Neural Network, and likely many other algorithms, it's a possibility that companies can rely on. By doing this project, I've gained a greater understanding of how to implement neural networks and how they work in tensorflow. Furthermore, I've experienced firsthand the deep data mining required to creatively feature engineer. Finally, I've better learned the six modules of MDS and how they relate data science to the scientific principles of the world. As I continue to pursue a career a machine learning, these principles of MDS will certainly be applicable across every project, especially any project related to scientific or mathematical principles.

Potential Areas to Improve

As this was my first major project in Mechanistic Data Science, there is a great deal to learn and improve on for the future. Beside the correlation matrix, my efforts to reduce the dimensionality of the data were not satisfactory due to the existence of binary and categorical variables in the data. In the future, I should do research on dimensionality reduction algorithms specifically catered to these types of variables. Potentially, this could also lead to a reduced order surrogate model, something I was unable to implement for the same reasons. While the neural network was a success, my unfamiliarity with tensorflow as a whole hindered my ability to fine tune the model and optimize it to its fullest potential. Furthermore, there are many other types of neural networks other than FNNs

which could produce better results, something to consider going forward. Overall, despite these shortcomings, they serve as future opportunities to learn, and do not take away from the machine learning skills I did learn while completing this project.

References

Liu, W. K., Z. Gan, and M. Fleming, *Mechanistic Data Science*. Springer Cham, 2021: p. 17-18. https://doi.org/10.1007/978-3-030-87832