DATA 100 Final Project (Traffic) Open-Ended Modeling Report

Group Members

- Shiqi Zhang (3035916454)
- Tina Chen (3035602660)
- Jerry Pan (3035788742)

Design Doc Clobber

- Motivation for feature
    - We believe the travel time differences between weekdays and weekends decrease after the lockdown. In other words, people take more time to travel in the same areas on weekdays compared to weekends before the lockdown and take less time to travel in the same area after the lockdown.
    - We also believe that Covid lockdown decreases people's travel time on both weekdays and weekends.

Problem

<u>Hypothesis Restated</u>

- We believe the travel time differences between weekdays and weekends decrease after the lockdown. In other words, people take more time to travel in the same areas on weekdays compared to weekends before the lockdown and take less time to travel in the same area after the lockdown.
- We also believe that Covid lockdown decreases people's travel time on both weekdays and weekends.

<u>Evaluation of the Hypothesis</u>

- The hypothesis can be evaluated by incorporating an extra feature of weekday/weekend identifier to train and predict the average speed. This can be found by observing whether incorporating the specific feature result in better model accuracy.

- By acquiring the model accuracy (r^2) in each model, we are able to confirm or reject the hypothesis. Improvement in model accuracy means incorporating specific features helps with better prediction and vice versa.
- The hypothesis can be confirmed or rejected with the provided dataset since the information we need are weekday/weekend identifiers and average speeds for weekday & weekend, pre-lockdown weekday & weekend, post-lockdown weekday & weekend. The first one can be acquired by looking at the calendar since we are only dealing with one month of data. The latter ones can be directly calculated by the provided dataset by taking the respective averages. As a result, this hypothesis can be evaluated assuming unlimited access to all data.

Answer

- The result shows that including weekday/weekend features indeed helps to better predict the overall average travel speed. Specifically, the weekday/weekend linear regression model improved its accuracy from 0.8972 as in the baseline model to 0.9057.
- However, the weekday/weekend pre-post lockdown model does not improve upon the weekday/weekend model with the same model accuracy of 0.9057.
- As a result, the result confirms the conclusion for the hypothesis. Incorporating the weekday/weekend feature indeed increases the model accuracy, yet separating this feature in pre lockdown and post lodown data does not contribute to a better model.

Modeling

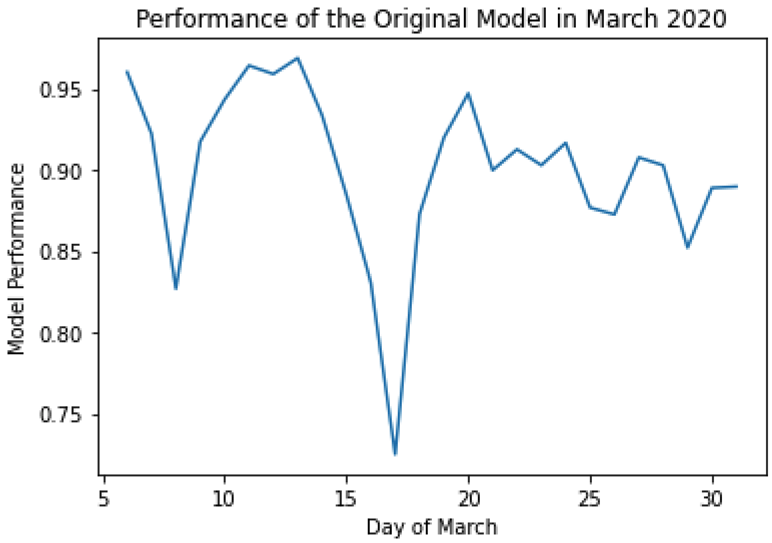
- We use linear model regression to train the average speeds of the census tracts across all days.
- The output of the model is to predict the average speeds before and after the lockdown. We use the average speeds of previous days to predict the average speeds of the target day through the model.
- Because we use average speeds and days in March as numerical data, we are able to build a linear model between these two variables. When we use linear regression as a baseline, even though most of the data fit under a linear model there are still some points, we saw there are some potential improvements towards the graph.

- By using feature engineering towards weekdays and weekends, we will combat the issue of underfitting. The feature engineering separates weekends and weekdays into categorical variables and takes the average speed on weekdays and weekends to increase the accuracy of the linear model. This solves our problems proposed on the EDA portion about the underfitting and better explains our hypothesis related to weekends and weekdays.

Model Evaluation and Analysis

- Supervised Learning Modeling
    - We predicted daily traffic speed per census tract given the previous k=5 daily traffic speeds for that census tract. In particular, say a matrix A is $n \times d$, where n is the number of census tracts and d is the number of days.
    - Inputs & labels
        - $X_{(i,t)} = [A_{(i,t-5)}, A_{(i,t-4)}, A_{(i,t-3)}, A_{(i,t-2)}, A_{(i,t-1)}]$
        - $y_{(i,t)} = [A_{(i,t)}]$
        - Simply put, each sample $X_{(i,t)}$ presents the speed average information for the previous 5 days for the ith census track
- Steps for Modeling
    - We firstly trained and evaluated the linear model reg = LinearRegression() on pre-lockdown data, achieving a R-square value of 0.9639. This result is really promising for the pre-lockdown prediction since there are not many changes regarding the traffic policy, political announcement, business & commercial activities, and etc., ending up in more consistent traffic patterns. Thus, we consider our first linear model to be really successful when generalizing over the pre-lockdown dataset.

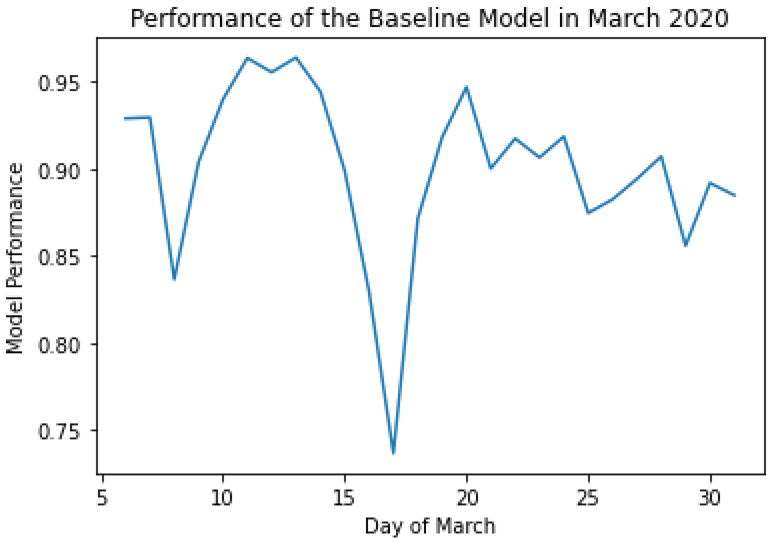
- As we applied our accurate linear model onto the post-lockdown data, there appears to be a number of anomaly predictions.
    - The model performance begins to worsen on March 15th, because all bars, nightclubs, wineries, and brewpubs are ordered to close, so the prediction just keeps getting worse from March 15th to 17th, where the model performance is the worst on the 17th. On March 16th and 17th, SFMTA announced cable car service will be suspended and Uber & Lyft suspended shared-ride options, which all significantly affect the traffic condition. Thus, this leads to the worst model performance on March 16th and March 17th.
    - The model miraculously recovers on its own as the data with changed policies will be reflected in the model, and thus improve the model prediction as aligned with the changed policies.



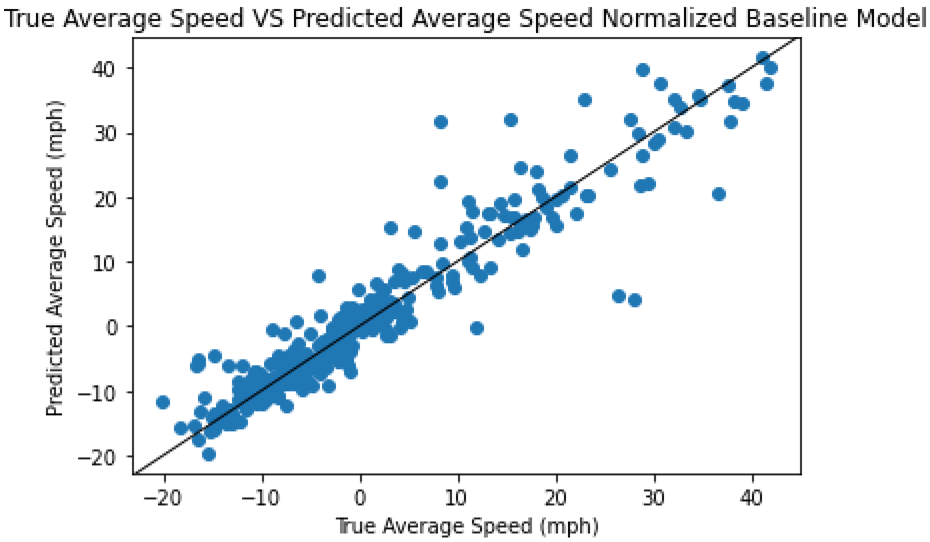
- To fix our linear model on post-lockdown data, we make the particular model improvements in the following section with the corresponding evaluation.

Model Improvement

- Main Approach for Improvement
    - Ensemble technique
    - Context-aware approaches
- Delta Naive Linear Model - Baseline model in Question 5a
    - Problem
        - The naive linear model doesn't work well on post-lockdown data because there are several days after the lockdown during which the traffic policy has undergone significantly so that the auto-regressive model doesn't work properly.
    - Solution
        - As inspired by ResNet (https://arxiv.org/abs/1512.03385), we believe this differential process would improve our feature engineering. To make the auto-regressive model accurate and reflect the changes in traffic policies, we normalize the speed entries.
        - After removing all NANs in the time-series dataset, we calculated the average speed across all days **ave_all_days** and normalized the original **time_series** to be **time_series_delta_baseline** by subtracting the **ave_all_days** from **time_series**.
    - Result
        - This simple technique helps us to achieve the model R-Squared value of 0.8972, which is already significantly better than our naive model as presented in part 4.
        - There are no concerns about overfitting the data because we intrinsically didn't manipulate nor transform the data in any ways.
        - Line plot: Performance of the Baseline model in March 2020



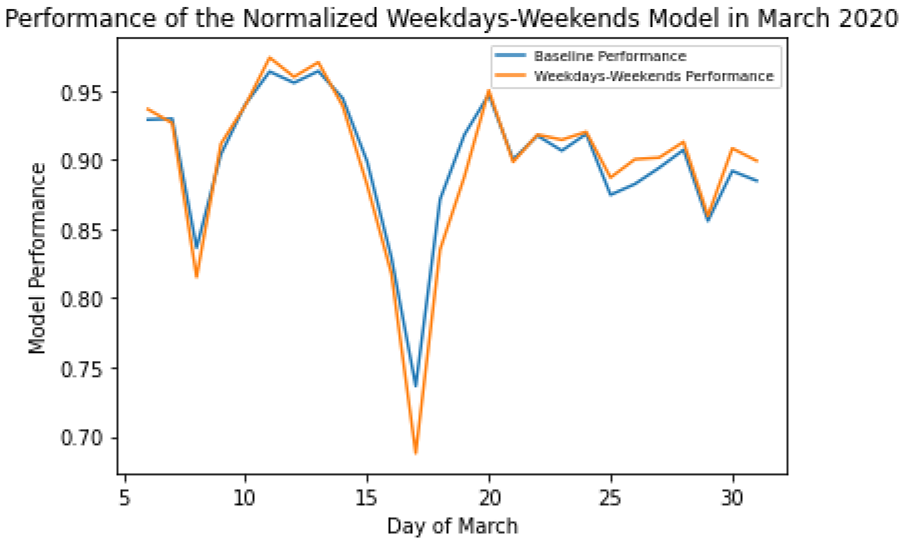
- The performance line plot shows approximately the same trend as in the original model with two dips model accuracies on day 8th and 17th. The baseline model shows the highest accuracy on day 12th and 14th. The accuracy has a general slight increase than the original model.
- Scatter plot: True Average Speed VS Predicted Average Speed Normalized Baseline Model



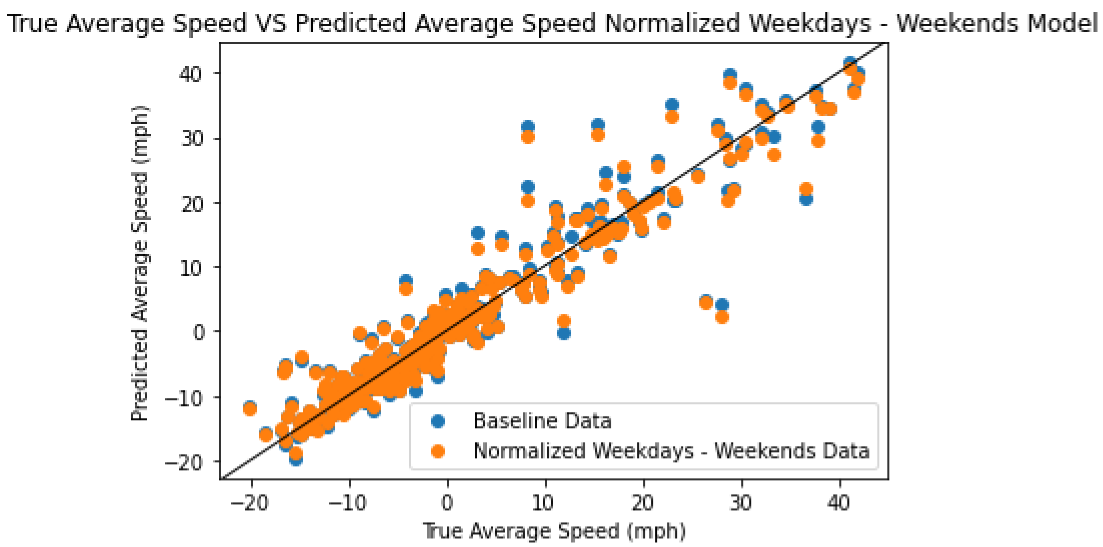
- The Scatter plot of the Baseline Model shows a positive linear relationship between the predicted average speed and treu average speed. Due to normalization, there are negative points in the plot. Those points do not indicate the true average speed or predicted average speed is negative. They are normalized to show the trend of the correlation between true average speed and predicted average speed. We can denormalize the average speed to obtain the original points. The Baseline model is already fitted with most of the data points. However, some points, especially when the true average speed is greater than 10 mph, do not fit the model well.

- Delta Weekend/Weekday Linear Model - Improved Model I in Question 5b
    - Problem
        - The Delta Naive Linear Model works pretty well, but the model could still be improved by partitioning up the weekday and weekend data so that we can perform delta differential normalization respectively.

- Solution
    - As inspired by decision trees, partitioning up data in certains ways could easily improve our linear model because we can specialize the model to fit each cohort of data more precisely. We further improved the model by partitioning up the weekday and weekend data so that we can perform delta differential normalization.
    - Though we proposed to partition up the weekday and weekend data, we still adopted one model. However, we calculated the speed average for weekday **speed_average_weekday** and weekend **speed_average_weekend**, and thereafter normalized the speed respectively into **delta_time_series_day_week**.
- Result
    - This simple technique helps us to achieve the model R-Squared value of 0.8972, which is already significantly better than our naive model as presented in part 4.
    - There are no concerns about overfitting the data because the weekend and weekday are treated separately. Besides, our previous open EDA indeed shows that the weekday and weekend display statistically significant differences in speed increased pre-lockdown vs post-lockdown.
    - Line plot: Performance of the Normalized Weekdays-Weekends Model in March 2020



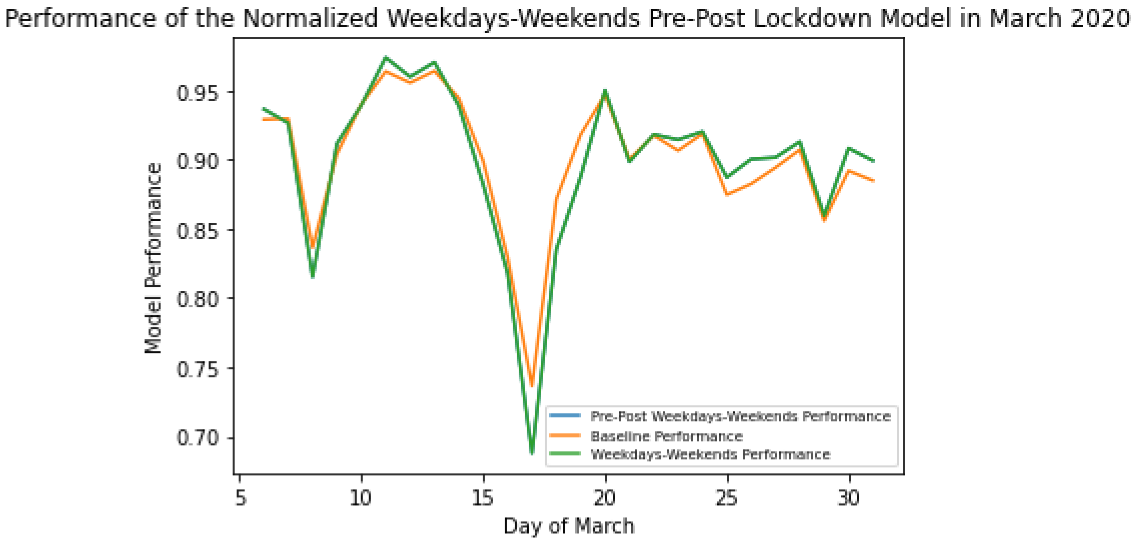
- This model shows the general trend of model accuracy as the baseline model. However, this model has higher accuracy on all points except for the two extreme dips on day 8th and 17th. The model results in worse prediction on these two days due to potential reasons for overfitting the data by incorporating new weekday/weekend features. The extreme result was exaggerated.
- Scatter plot: True Average Speed VS Predicted Average Speed Normalized Weekdays - Weekends Model



- We improved the model by separating the average speed by weekdays and weekends and overlaid the scatterplot of the new model on the baseline model. Due to normalization, there are negative points in the plot. Those points do not indicate the true average speed or predicted average speed is negative. They are normalized to show the trend of the correlation between true average speed and predicted average speed. We can denormalize the average speed to obtain the original points. From the graph, we can see the improvement from the baseline data to the new normalized data. The improvement is especially obvious as the true average speed increases. In order to further improve the data at the right upper corner, we decide to further separate the data into pre lockdown and post lockdown.

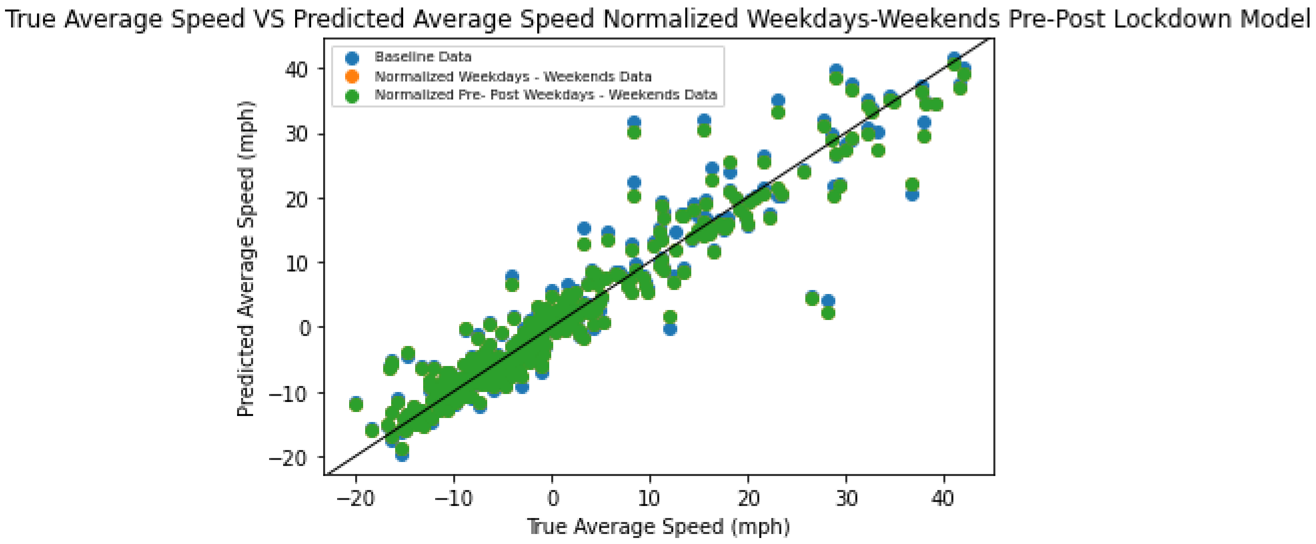
- Delta Weekend/Weekday pre/post-lockdown Linear model - Improved Model II in Question 5c
  - Problem
    - Technically, there aren't any obvious problems with our previous Delta Weekend/Weekday Linear Model, but we still want to further optimize it as perfectionists.

    - To be particular, the lowest point in model performance is still under 0.75, which is even lower than our previous naive model in Q5a, motivating us to fix this part.
- Solution
    - In the design of decision trees, we further partitioned data in for even more data normalization so that the model could fit each cohort of data more precisely. We further improved the model by partitioning up the pre/post-lockdown weekday and pre/post-lockdown weekend data so that we can perform delta differential normalization.
    - In this case, we calculate the speed average for pre/post-lockdown weekday and pre/post-lockdown weekend data respectively, i.e. **speed_average_weekend_pre**, **speed_average_weekend_post**, **speed_average_weekday_pre**, and **speed_average_weekday_post**.
    - Furthermore, we filter out the outlier speed average data (not within mean +- 2std) to increase model generalization.
- Result
    - There are no numerical improvements in model R-Squared value. We speculated that overfitting could happen since there are only 31day of traffic data, so our performance is already really good.
    - If given more data, we believe that our model would be a lot more robust and accurate.
    - Performance of the Normalized Weekdays-Weekends Pre-Post Lockdown Model in March 2020



- The Pre-Post Weekdays-Weekends model overlapped with the Weekdays-Weekends model entirely. This model shows the general trend of model accuracy as the baseline model, having higher accuracy on all points except for the two extreme dips on day 8th and 17th. The model results in worse prediction on these two days due to potential reasons for overfitting the data by incorporating new weekday/weekend features. However, the split in pre lockdown and postlockdown does not have any impact on the model performance.

- Scatter plot: True Average Speed VS Predicted Average Speed Normalized Weekdays-Weekends Pre-Post Lockdown Model



-

- We further improved the model by separating the average speed by pre lockdown and post lockdown and overlaid the scatterplot of the previous models on the new graph. Due to normalization, there are negative points in the plot. Those points do not indicate the true average speed or predicted average speed is negative. They are normalized to show the trend of the correlation between true average speed and predicted average speed. We can denormalize the average speed to obtain the original points. From the graph, we can see the improvement based on the new normalized data is trivial to observe. Since there are only 31 days of data, the second improvement model is overfitted. This offers us directions to further improve the model.

Future Work

- In the real world, the dataset is always involved with complex uncertainty. While we've been pushing hard to achieve higher model validation accuracy, we believe the data-centric model would be more beneficial by gleaning more data instead of simply working on bettering our models. Thus, if possible, we can collect and clean more data to better train our models while avoiding the overfitting, where our current model would be improved and be precautioned on.
- Since our second improved model has the issue of overfitting, our future work can be joining external data sets such as speed average after March 2020 or speed average by time slot to increase our data set. Joining the new data set will help us better understand and improve the model since more detailed data entry with small granularity would help our feature engineering and thus make the model with more information.