2018 TAMIDS Data Science Competition
Team: Bulgogi

Tae Jun Jeon
Texas A&M University
Department of Computer Science and Engineering
tjeon90@tamu.edu

Donghwa Shin
Texas A&M University
Department of Computer Science and Engineering
donghwa_shin@tamu.edu

Abstract

The 2018 TAMIDS Data Science Competition is geared towards understanding the taxi business of Chicago based on the trips associated during the 2013 to 2016. The challenge itself is more suited for feature engineering where performance of the regression of predicting the median revenue is dependent on including different features. What we propose is very different from that of the status quo. Our model takes into account the seasonal and trends of the taxi business forecasting the number of rides for the target year. We train a neural network to calculate the median weekly revenue based on the total revenue earned for each route. We use the obtained values from our forecasting method as inputs for the neural network to predict the median weekly revenue.

We believe that our work is noble because our approach in handling uncertainty allows the model to make excellent predictions. Our model outperformed many of the finalists in terms of making predictions for the median weekly revenues for 2017. Further insights will be discussed during the result section.

1. Introduction

The primary task for the 2018 TAMIDS Data Science Competition is to be able to explain the change in the taxi industry in Chicago by providing visualizations and predictive models were given the dataset of 2013 to 2016. We were asked to provide the hourly, daily and weekly revenue of a typical Chicago taxi based on the change in both location and time.

Revenue of a typical taxi driver is dependent on the total number of rides and how much each of the rides cost. Hence, we have performed a preliminary analysis of the dataset to understand what causes the number of rides and total fares to change. We have visualized the data by changing the parameters to see how much the total rides were influenced by time and location. We then observed the impact of different features on the total fare of individual rides. After observing our results, we have concluded that the median revenue will be impacted more on the forecast of the number of rides.

Our contributions are summarized below:

- We make a preliminary analysis on the dataset making a distinction of how routes of a ride react differently to change of time.
- We perform forecasting on routes for the given time frame of weeks to make predictions of number of rides for routes in weeks in 2017.
- We sample the fore-casted number of rides from training data to get an estimate of total fare cost of each route
- We train the neural network with weekly total revenue per route and weekly median revenue.
- We use the predicted weekly total revenue per route as an input to obtain the weekly median revenues for our target weeks in year 2017.
- We will discuss the results of our model and the limitations of our current approach.

2. Dataset

The City of Chicago has released Chicago Taxi Data containing more than 110 million Chicago taxi rides from Jan. 1st, 2013 to Jul. 31st, 2017[8], and we are given this dataset for the competition. The dataset has 23 features including 'Taxi ID', 'Trip Seconds', 'Trip miles', 'Pickup/Dropoff locations', 'Total Cost', etc. Table 1 shows the features we have. Also, We are asked to use the data from 2013 to 2016 for training and the data from 2017 for testing purpose.

Trip ID	**Taxi ID**
Trip Start Timestamp	Trip End Timestamp
Trip Seconds	Trip Miles
Pickup Census Tract	Dropoff Census Tract
Pickup Community Area	**Dropoff Community Area**
Fare	Tips
Tolls	Extras
Trip Total	Payment Type
Company	Pickup Centroid Latitude
Pickup Centroid Longitude	Pickup Centroid Location
Dropoff Centroid Latitude	Dropoff Centroid Longitude
Dropoff Centroid Location	-

Table 1: 23 features that the dataset contains. We only used bold-faced features for our framework

The dataset is very messy, which means taxi trips in the dataset have changed due to a privacy issue. For example, a pickup location of each taxi trip is not the exact location that a taxi driver picked up his/her customer, but the general area. On top of that, the dataset contains some weird taxi rides information such as a taxi ride whose cost is $1,000 and trip miles is 0, which does not make sense. We intend not to remove these data from the dataset and in our experiments these data are used without any modification. It also means these data are included when we compute our ground truth, which is the actual weekly revenue for each week from 2013 to 2017.

3. Related Work

Time series forecasting is used widely in the research areas of economics, finance and weather predictions[5]. The idea behind time series is to grasp the trends, seasonality and residual signals in making predictions for upcoming independent time periods[2]. There are mainly two methods of observing trend. They are moving average and exponential smoothing[2]. The moving average observes previous k iterations equally in calculating the average whereas the exponential smoothing values the most recent iterations more. The seasonality can be obtained by using the trend values to divide it to the training data to see how much is being multiplied for that particular time frame[2].

There also have been a lot of studies on improving taxi business models using advanced technologies such as the global positioning system (GPS)[9, 7, 6]. Mukai et al.[6] showed that neural networks can be used to forecast the taxi demands based on a lot of customer trajectory data.

Recently, Cramer et al.[4] studied how taxi business has been affected by ride sharing services such as Uber and Lyft. The study illustrated that the Internet-based technologies which help matching passengers are effectively providing competition in the taxi industry.

4. Preliminary Analysis

During the preliminary stage of the competition, we wanted to understand the taxi data. Mainly we wanted to observe how the trips were spread out and if anomalies could be observed by spreading the data by time and location.

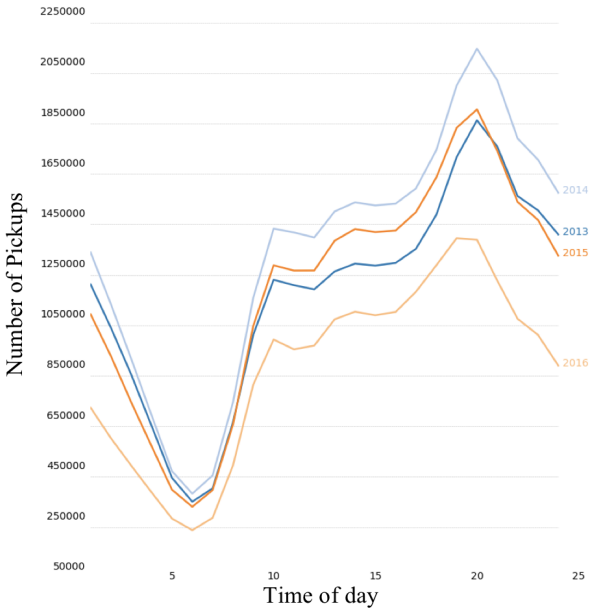

Figure 1: Number of rides per hour

Figure 1 shows the total number of rides placed into the respected hour of the day. We can see that the general pattern stays the same for all years. We can observe that there was an increase from the year 2013 and 2014. It is observed that there is an accelerated number of decreased as it gets closer to 2017.

Figure 2 shows the total number of rides placed into the respected month. While the patterns do not match as well as that of Figure 1, we can still make an observation that certain months tend to be higher than others for the same year. Also the pattern of a dip existing at April was not observable for year 2016.

Figure 3 shows the total number of rides placed into the starting location where the customers were picked up. We can see that most rides were requested in small number of locations and the decreasing pattern of rides is evident for all locations.

Figure 4 shows the median revenue for a typical taxis driver per week. We were able to make an observation that patterns for 2013 to 2015 tend to be similar whereas 2016 had some irregular patterns. There were four big dips that were not observed in the previous years.

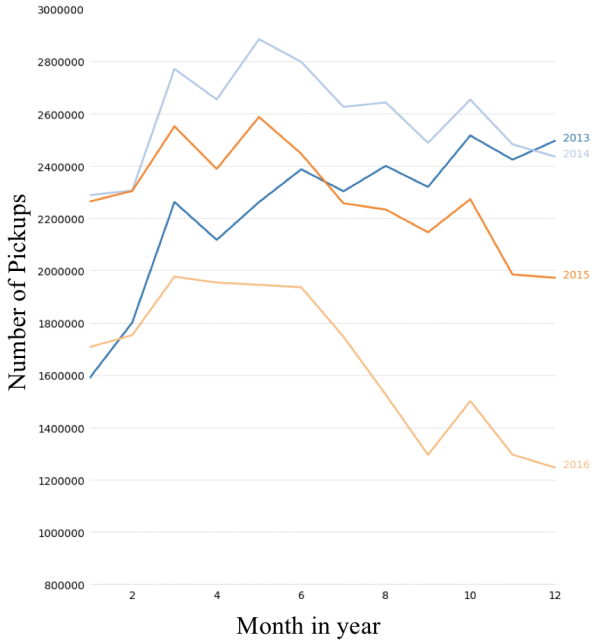

Figure 2: Number of rides per month

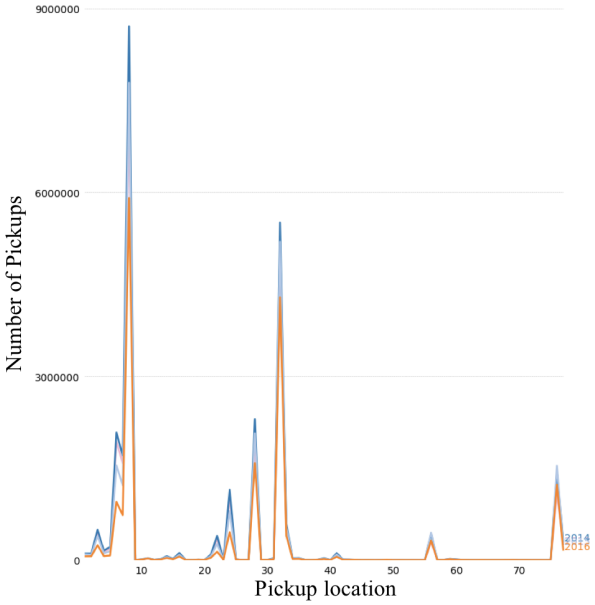

Figure 3: Number of rides per location

From observing the four figures above, we concluded that there are some anomalies in the patterns for the 2016 dataset. In general, we could not find any abnormal behavior where certain locations or time periods did not follow the general trend of decrease of taxi rides. Hence, we decided

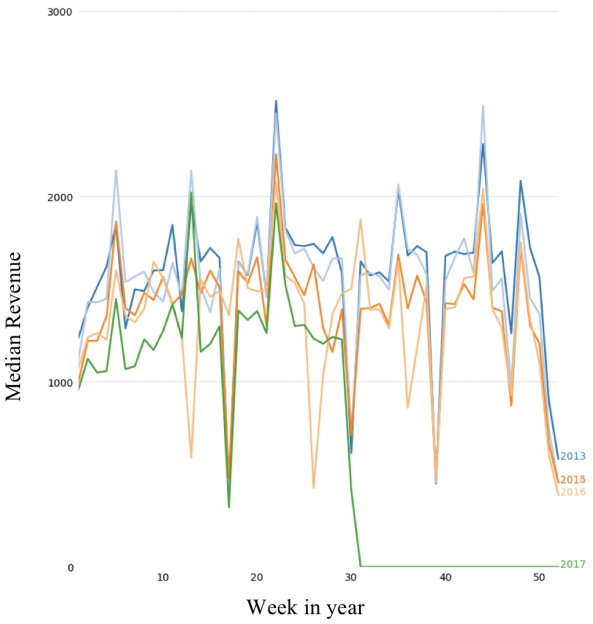

Figure 4: Median revenue per week

not to perform clustering algorithms to capture abnormalities and proceed with our following model.

5. Prediction Model

The median revenue prediction depends on the number of trips and the cost of each individual trip. The problem is that each trip may cost different depending on the route, the length of travel, type of payment, and etc. So we came up with an intuitive way that uses both forecasting methods and neural network to produce an intuitive outcome.

Figure 5 shows the general flow of our model. We initially divide the rides into weeks and count the number of transactions for each route, pickup to drop off. Then we perform forecasting algorithm on the weekly divided individual routes and predict how many routes will be requested for the 52 weeks in 2017. This number of routes will be used in randomly sampling trips from previous years to get the total fare and revenue will be calculated for each route for each week. With the resulting matrix we use it as an input for the neural network that was previously trained to calculate median revenues based on total revenue per route.

In our framework, we have 5 phases, and explain each phase of the framework in detail.

5.1. Extract features

In the first phase, we extract features from the given training datasets. The dataset of each year is divided into 52 chunks of data since a year has 52 weeks. After that, rides

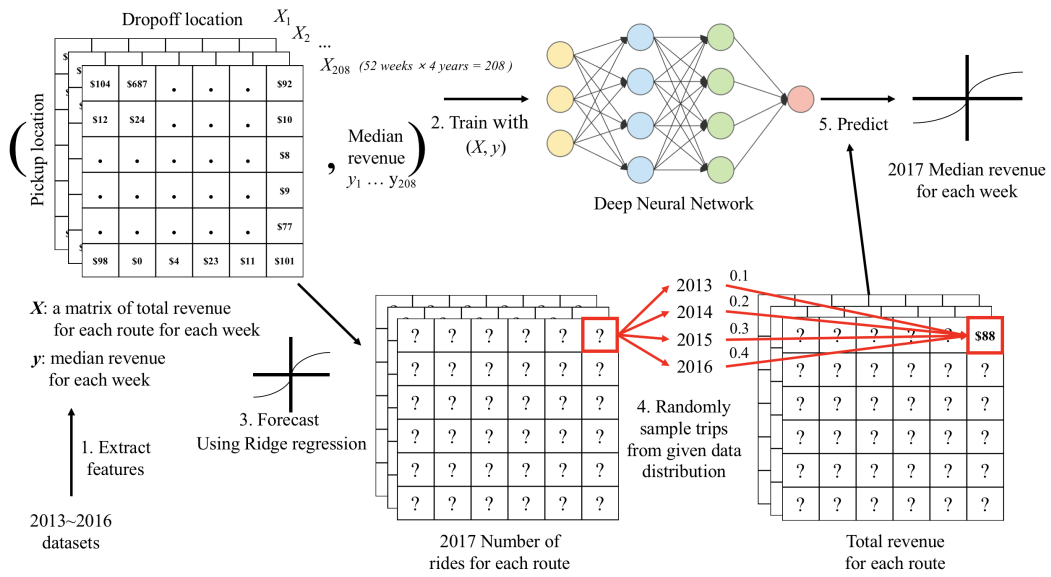

Figure 5: The Proposed Framework

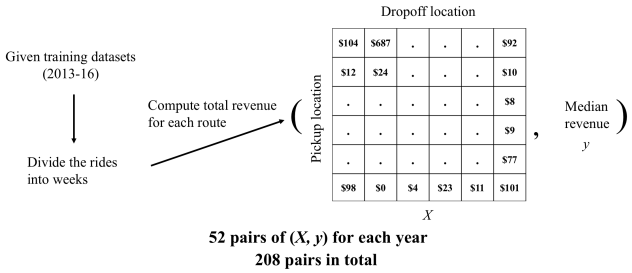

Figure 6: Extract features

Layers	# of neurons
Input layer	5929
Hidden layer 1	512
Hidden layer 2	256
Hidden layer 3	128
Hidden layer 4	256
Hidden layer 5	512
Hidden layer 6	5929
Output layer	1

Table 2: Neural network regressor experimental setting.

of each week are mapped into a 77×77 matrix X where each element represents total weekly revenue corresponding to each pickup location and dropoff location. Also, we compute median weekly revenue y for each week and it is used as our ground truth for the framework. As a result, we can get 52 pairs of (X, y) for each year and 208 pairs in total.

5.2. Train a deep neural network for regression

With 208 pairs of (X, y) that we have created in the first phase, we train a deep neural network regressor. For this, we use Keras[3] and TensorFlow[1] in the backend. The neural network setting is shown in Table 2.

The neural network has 8 layers including 6 hidden layers, an input layer, and an output layer and the number of neurons for each layer is 5929, 512, 256, 128, 256, 512, 5929, and 1, respectively. Rectified Linear Units (ReLU) are used for the activation functions for all of the layers.

5.3. Forecast the number of rides using linear regression

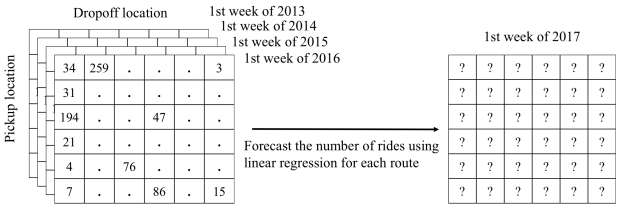

Figure 7: Forecast the number of rides

In the third phase, we forecast the number of rides using simple linear regression models such as Lasso and Ridge. In our experiments, our baseline method with Ridge regression is used for this task. How we implement our baselines with Lasso and Ridge is introduced in Experiments section. By

using this method, for example, we can forecast the number of rides for each route for the 1st week of 2017 based on the number of rides for each route for the 1st week of 2013 to 2016. Figure 7 illustrates this procedure.

5.4. Randomly sample rides from the training datasets

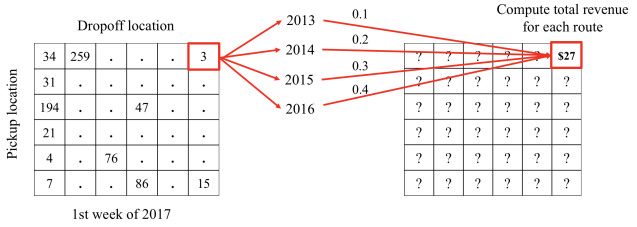

Figure 8: Randomly sample rides from the training datasets

After forecasting the number of rides for 2017, we sample rides from the datasets from 2013 to 2016. For example, if the estimated the number of rides is 3, then we sample 3 rides from the datasets based on the probabilities we set to each year's dataset, which are 0.1, 0.2, 0.3, and, 0.4, respectively. The rationale behind these probabilities is that we believe the more recent data, the more relevant. After sampling rides, we simply compute total revenue for each route based on the sampled rides.

5.5. Predict median weekly revenue

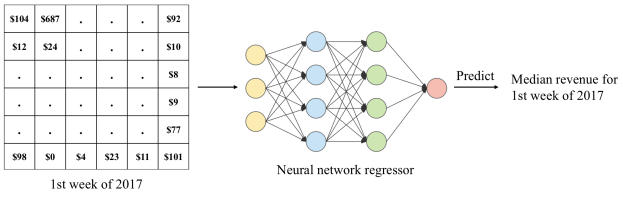

Figure 9: Predict median weekly revenue

After the 4th phase, we have a 77×77 matrix X for each week of 2017, which is the same type of matrix that we use for training the neural network regressor. Now, we put it into the trained neural network and predict median weekly revenue for each week of 2017.

6. Experiments

6.1. Baselines

The following is the breakdown of the regression process needed for forecasting the number of rides for each route for the target year. Also, this method is used for our baselines.

6.1.1 Trend

The trend represents the general movement of the taxi business. We chose to use the moving average method. The moving average can be obtained by taking the average of previous k number of weeks. We set the window size, k, to 52 which represents the number of weeks in a year.

6.1.2 Seasonality

The seasonality can be best explained as depicting repeated patterns. The seasonality is obtained by taking the average of the change calculated for each i-th week from the different years in the training set. The change can be obtained by dividing the total ride of the week by the moving average.

6.1.3 Regression

We use Lasso and Ridge regressions to propagate the moving average values onto the target year. The prediction for the target year will be obtained by multiplying the seasonal values with the regression values for the weeks in 2017. This way the general trend of the taxi business and its seasonal effects can be incorporated. For the baselines, both Lasso and Ridge are used.

6.2. Evaluation

We use RMSE (Root Mean Squared Error) as our evaluation metric on the predicted median weekly revenue values. RMSE values are calculated as follows.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(a_i - p_i\Big)^2} \quad (1)$$

where a_i and p_i denote actual weekly revenue and predicted weekly revenue for i-th week of 2017, respectively, and n is the number of weeks in a year, which is 52.

6.3. Experiment Results

We compare our model with the baselines mentioned above. Figure 10 to 13 show the experiment results. Due to the anomalies detected in year 2016, we perform the baseline models along with our model with two datasets. One that uses all the data from the training data and one excluding 2016. Table 3 shows our experiment results with RMSE values for the baseline models and our model.

RMSE	Lasso	Ridge	Our Model
2013-2016	194.7246	194.7246	**159.8496**
2013-2015	237.2911	237.2911	**108.3683**

Table 3: Experiment results

As you see the table, the proposed model outperforms the baselines with different datasets, 2013 to 2016 and 2013

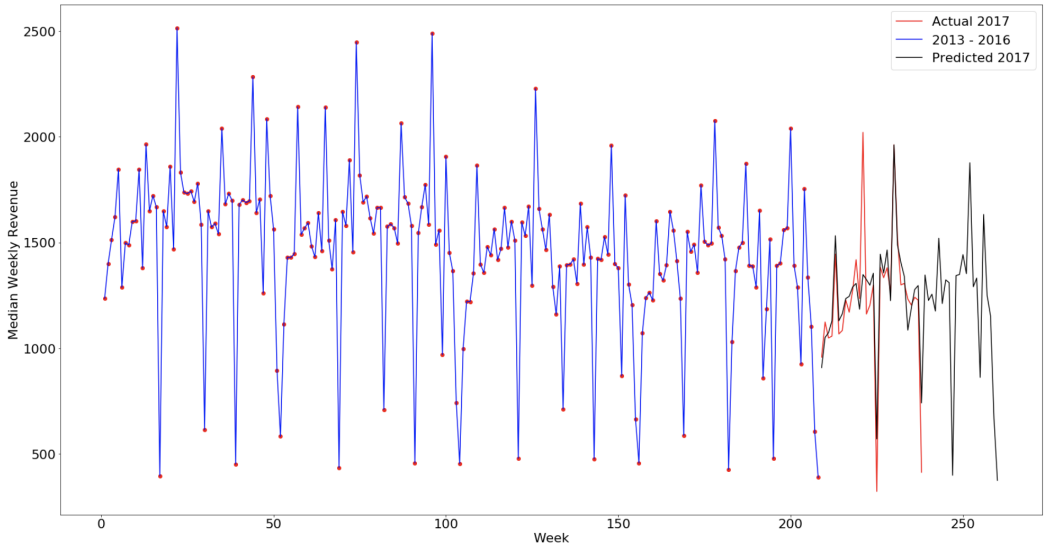

Figure 10: Median weekly revenue for 2013 to 2016 and predicted median weekly revenue for 2017 by our model based on the given data from 2013 to 2016. The red dots in the figure represent median weekly revenues for each week, the red line represent the actual median weekly revenue, and the black line represents the predicted median weekly revenue by our framework.

to 2015. The two baselines show exactly the same results. Interestingly, the performances of the baselines get worse with the dataset from 2013 to 2015 than from 2013 to 2016. However, our framework with the dataset from 2013 to 2015 shows better performance as we expected.

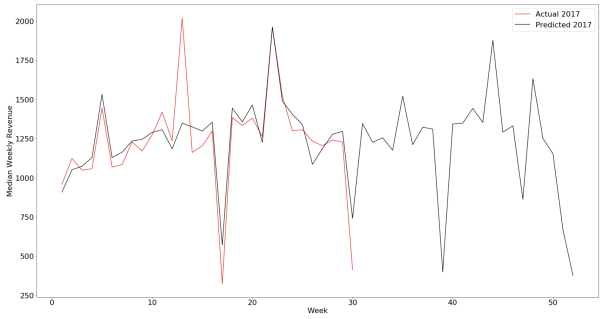

Figure 11: Predicted median weekly revenue for 2017 by our model based on the given data from 2013 to 2016

Figure 11and 12 illustrate the median revenues predicted by our model and Ridge regression, respectively, for each week of 2017. As you see the figures, our model successfully predict the median revenues for the first 30 weeks except the period from week 10 to 15. We speculate that it is derived from the drastic decrease of the median value for 2016 in that time frame.

In order to figure out the cause, we also conducted an experiment with the dataset from 2013 to 2015 and Figure 13 shows the predicted median weekly revenues for 2017 by our model. The result says the framework could predict better than with all of the datasets as we speculated.

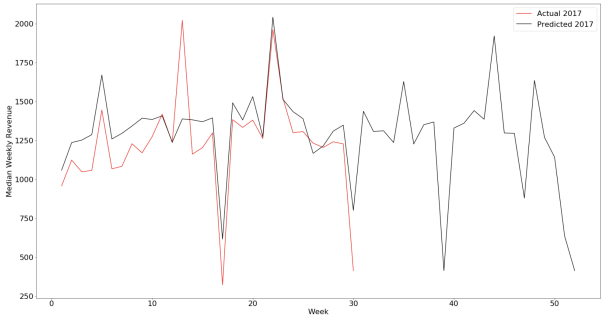

Figure 12: Predicted median weekly revenue for 2017 by Ridge regression based on the given data from 2013 to 2016

7. The Competition

This section is devoted to explain the other projects of the competition. As explained in this project report, we improvised a model that can observe the bare minimum features of the taxi ride information to make reasonable prediction based on machine learning and forecasting. The other finalists focused on creating new features to strengthen the

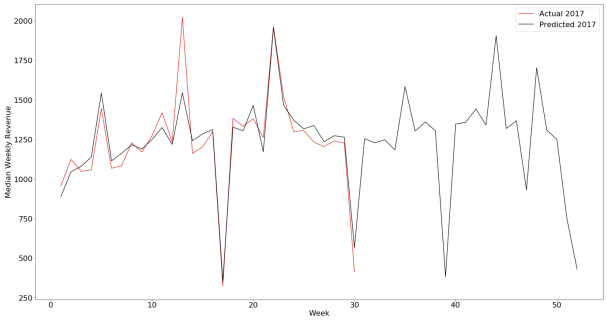

Figure 13: Predicted median weekly revenue for 2017 by our model based on the given data from 2013 to 2015

linear predictions. Many teams who had more teammates individually divided roles to address the prediction task and visualization tasks. Notable works were performed by Stat graduate students with feature engineering experience. This proved to be useful, considering abnormal behaviors were caused by special events such as the St. Patrick's parade in 2016 week 10 to 15, observe the orange line drop from figure 4. Our model did not address this behavior, which led to a lower prediction for that week, observe figure 11. Our model has a strong assumption that the trends and seasonality will be stable. From our dataset, that was not the case. While we were not able to make it to the top 3 who received awards from the committee, we believe that our novel approach provided a good reputation and capability that the computer science can bring to the table of data analytics.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning.

[2] C. Chatfield. *Time-series forecasting*. CRC Press, 2000.

[3] F. Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[4] J. Cramer and A. B. Krueger. Disruptive change in the taxi business: The case of uber. *American Economic Review*, 106(5):177–82, 2016.

[5] E. S. Gardner. Exponential smoothing: The state of the art. *Journal of forecasting*, 4(1):1–28, 1985.

[6] N. Mukai and N. Yoden. Taxi demand forecasting based on taxi probe data by neural network. In *Intelligent Interactive Multimedia: Systems and Services*, pages 589–597. Springer, 2012.

[7] K. Tamai and A. Shinagawa. Platform for location-based services. *Fujitsu Sci. Tech. J*, 47(4):426–433, 2011.

[8] The_City_of_Chicago. Chicago taxi data released. https://digital.cityofchicago.org/index.php/chicago-taxi-data-released/, 2017.

[9] C. Wang, W. K. Ng, and H. Chen. From data to knowledge to action: A taxi business intelligence system. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 1623–1628. IEEE, 2012.