To: Paul Alley (CoS IT) & Mary Catherine Snyder (SDOT)
From: UW Team - Shreya, Sahil, Allison, Nathan
Date: May 15, 2019
Subject: Summary of Parking Model Results

This memo summarizes the results of model development for predicting on-street paid parking occupancy rate for Belltown North. Comprehensive documentation about process and results will be provided prior to the completion of the project.

General process

We split the data into training, validation and test sets. Because this is temporal data, these splits were made based on date. Training and validation sets were used for model development and the test set was only used for determining final model performance.

A variety of models were explored. Each model was trained using the training set. Predictions were made using the validation set and performance metrics were computed based on the accuracy of those predictions. The performance metrics used were R-squared, adjusted R-squared, and root mean squared error (RSME). For models with promising initial results, we tuned hyperparameters to achieve the best prediction performance on the test set.

Our final model was chosen based on the performance on the validation set. The model was then retrained using the combined training and validation sets and performance was measured on the test set.

Feature engineering and time series features

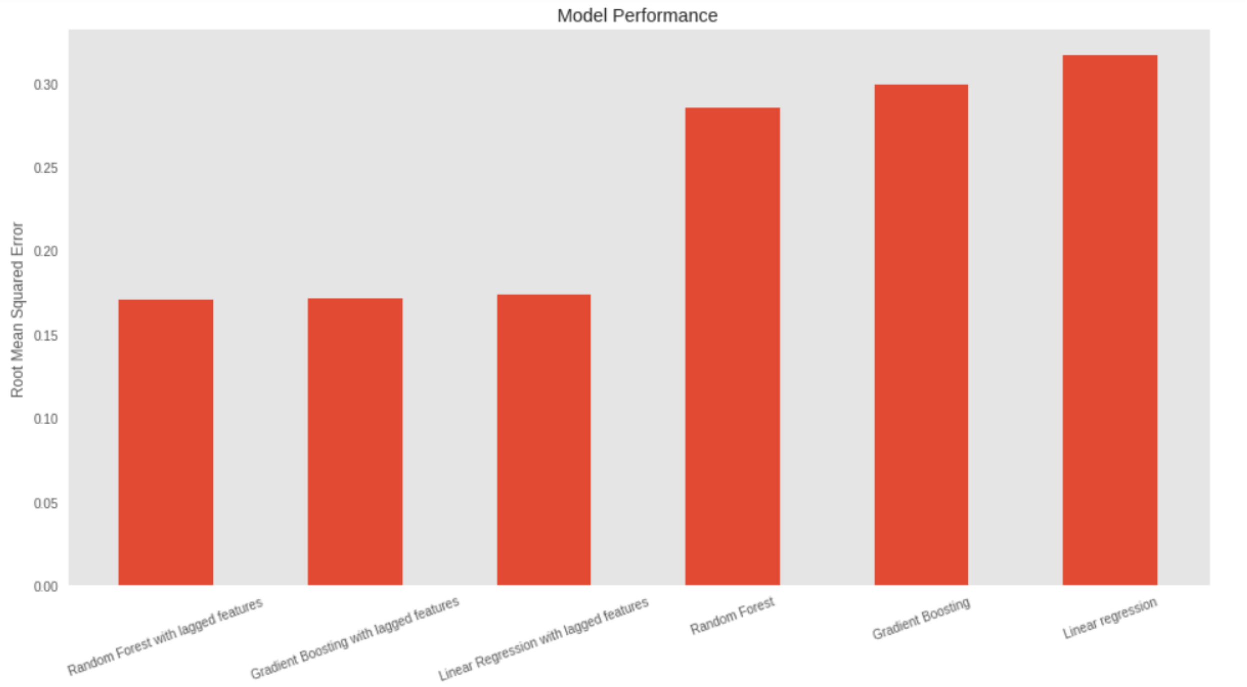
We developed a number of features in an attempt to better capture factors related to parking occupancy. These include:

- Count of business licenses, split by category, within 300 meters (approximately 2 blocks)
- Relative distance from downtown
- Relative distance from the waterfront
- Whether the street is a principal or minor arterial

Additionally, because parking in one period depends on parking in the previous period, we included a number of time series features. First, we added features like the cosine and sine of the hour. These features help to more accurately capture the cyclic nature of these variables. Additionally, we added lagged features, like the occupancy in the previous half hour, the previous day, and the same period in the previous day.

Evaluation of models

Based on preliminary results, we evaluated linear regression, random forest, and extreme gradient boosting machines. We evaluated models both with and without lagged occupancy features. The RMSE of the best model in each category is shown below; lower RMSE is better.



Models with lagged features clearly performed better than models without lagged features. Parking occupancy in one period depends on occupancy in the previous period and these variables are crucial for prediction performance. If prediction is not the goal, other models could be used.

The three models with lagged features performed very similarly. Linear regression has the benefit of being a simple and highly interpretable model, however, random forest performed best without lagged features as well. This leads us to believe that random forest is better capturing the factors related to parking occupancy that are of interest to the City of Seattle. Additionally, linear regression residual plots showed potential problems with the model. Therefore, we chose random forest with lagged features as our final model.

Final model performance

As explained above, we re-trained the random forest model on the combination of training and validation sets and then used that model to predict occupancy using the test set. Our final RMSE was 0.17, very close to the RMSE obtained in model development. This means that on average, the model's prediction vary 17 percentage points from actual occupancy. Given the complexity of the data and parking behavior, we believe this is a fairly good result, though additional data could potentially help to boost prediction accuracy.

Relevant factors

The following features emerged as the most relevant to parking occupancy in the final random forest model:

1. Occupancy 30 minutes prior
2. Occupancy 60 minutes prior
3. Occupancy in the same period on the previous day
4. Sine of hour
5. Cosine of hour
6. Relative distance from downtown
7. Count of medical offices within 300 meters
8. Total count of all businesses within 300 meters
9. Count of bars and restaurants within 300 meters
10. Count of grocery stores and markets within 300 meters