Predicting Weekly 2020-2024 Boston BlueBike Ride Count

Sophie Hwang, Data Science Capstone Project

Background

- Boston's BlueBike public bike share program is an essential affordable, healthy means of transportation with over 27 million trips since its launch in 2011 as Hubway.
- By analyzing the relationship between total BlueBike weekly trips and the relevant weather, transportation and social information, we can improve the business model and the experience of users.
- **Research Question: What variables help predict Boston BlueBike weekly ridership between 2020-2024?**

Data and Methods

Data Acquisition:

1) Monthly Boston BlueBike Data:
   Variable: Individual Trip Data for all BlueBike trips taken 2020-2024 Continuous
2) Boston Weather Data (retrieved via Meteostat API):
   Variables:
   - Daily Temperature (Celcius) Continuous
   - Daily Precipitation (millimeters) Continuous
   - Daily Wind Speed (km/hr) Continuous
3) MA Gas Price: Weekly Average Gas Price (USD) Continuous
4) Boston COVID-19 Cases Data: Daily New Cases Continuous

Data Cleaning:

1) Merge BlueBike Data to calculate the weekly total trip count in selected date range
2) Calculate weekly average for weather, gas, and COVID datasets and merge with the BlueBike dataset
3) Impute missing values with 0 - the only missing values upon inspection were early 2020 COVID case count
4) Finalize a working dataset with *259 rows & 8 features*.

Therefore, this data is only generalizable for weekly ride totals within the selected date range of 2020-2024.

Methods:

1) Linear Regression
   - First Order: AIC & BIC Comparison based on stepwise regression
   - Interaction: AIC & BIC Comparison based on stepwise regression
2) Support Vector Regression:
3) **K-fold Cross Validation**: equally split data into 10 folds using designated indexes; For each model, calculated the mean CV Score.
4) Final Model Selection & Fitting: Fit the entire dataset on the chosen model (BIC Interaction model)

Model Comparison

	Adjusted R^2	RMSE	AIC	BIC	F-statistic	p-value
First Order	0.7344	18479.77	5889.063	5921.178	NA	NA
Interaction - AIC	0.8396	15067.32	5770.875	5856.515	12.117	0
Interaction - BIC	0.8329	14890.25	5778.079	5849.446	14.608	0
SVR	NA	20014.7	277 Support Vectors			

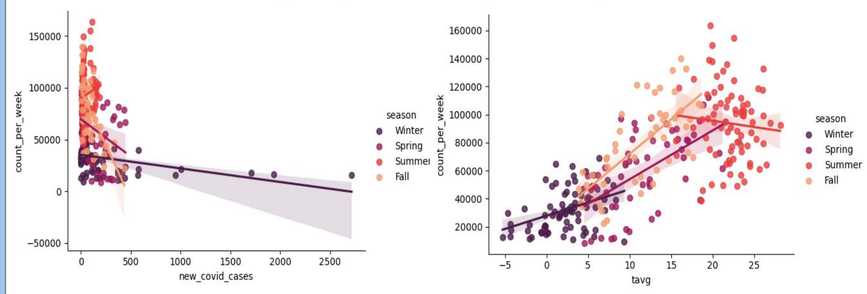
Table 1. Comparison of different linear regression models and support vector regression. Partial F-test and p-values are each compared to the first order model.

Final Model

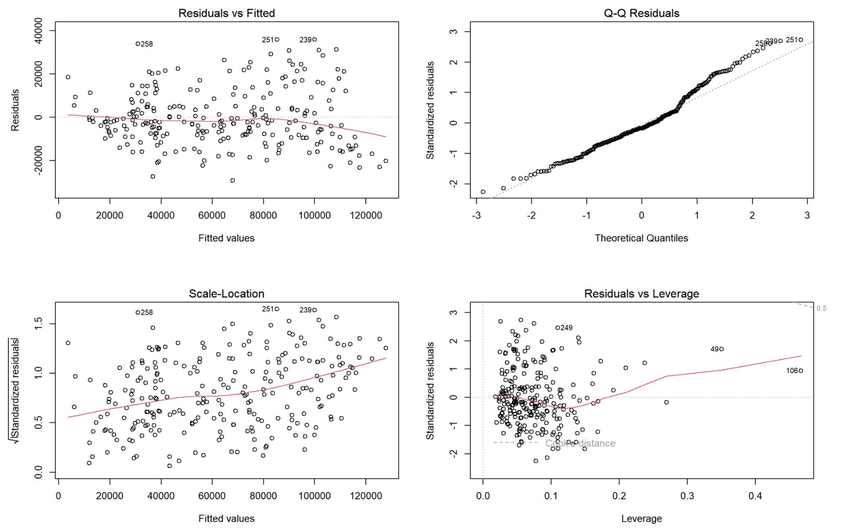
	gas_price	new_covid_cases	tavg	prcp	month:Summer	covid:SUmmer	covid:Winter	tavg:Spring	tavg:Summer	tavg:Winter	gas:tavg	Adj. R^2
Coefficients	14446.58	-70.07	2795.91	-897.39	14728.25	-100.09	59.29	-2597.86	-4544.52	-2691.82	450.48	0.8413

Table 2. Final model based on interactive AIC. The table only includes significant predictors.

New covid cases & avg temp vs. bike trip count by Season



Residual Plots



Conclusion

According to the model comparison table, the best model is AIC Model with interaction which indicates that there is a correlation between *Weekly BlueBike trips* and *Weather, Gas Price, COVID cases, and seasons.*

- We use the Adjusted R^2 over RMSE because we prioritize a model that accounts for the most variation in the data. Though BIC Interaction has a lower RMSE, it is a difference of ~70 rides per week.
- The R^2 indicates **84.13%** of the variation in weekly BlueBike ride count could be explained by the predictors and their interactions.
- Cross validation shows that the model is not overfitting.

Discussion

1) Normality & Equal Variance Assumption: As seen in the residual plots, the assumptions are satisfied.

2) Time Correlation: We have made the assumption that the weekly aggregate counts in our data set were independent of each other, not taking into account correlation between time and adjacent dates.

3) Future Direction: Conduct time-series relevant analysis on weekly ride data. We could also conduct similar analysis on other public transportation data and compare models and results.

References

1. Boston BlueBike Data https://bluebikes.com/system-data
2. Boston Weather Data https://dev.meteostat.net/python/#installation
3. MA Gas Data
4. Boston COVID-19 Dashboard