Can Search Predict TV Ratings?

mattmates@

Project Proposal

Background: Historically we have seen very high correlations between search volume and business KPIs. Search data is currently being used across Media & Entertainment to help inform and predict the sales of video games, event tickets, and movie box-office.

Problem: Television ratings have been on the decline for nearly a decade and TV companies like Disney, NBC Universal, ViacomCBS, Discovery, etc. are looking for earlier indicators to understand if their next new TV show will meeting their ratings goals & decide how to allocate marketing budgets

Hypothesis: Google Search is the world's best measure of interest and intent and can be a strong predictor of how well a TV show will premiere

Goal: To understand the relationship between Google Search data and TV shows then determine if Google Search data is a strong predictor of TV Viewership

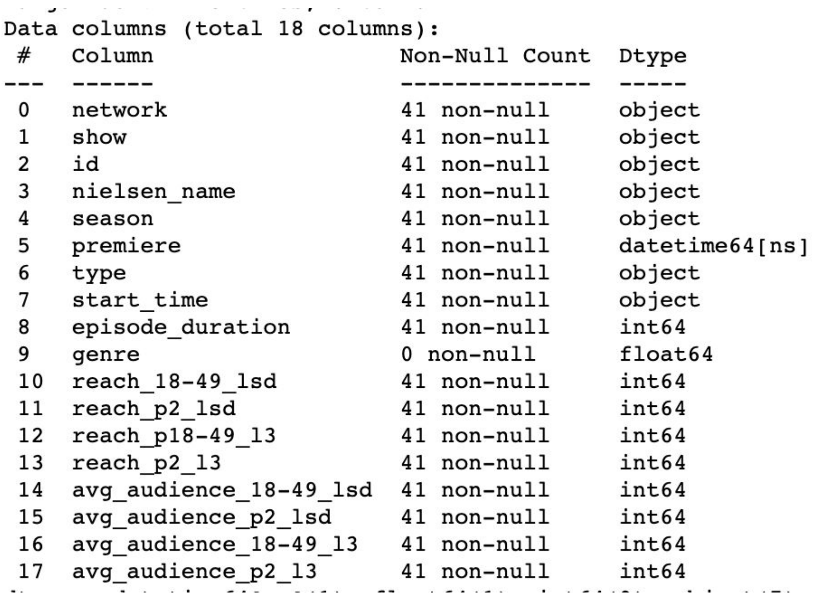
Success Metrics: Model Accuracy

Risks/Limitations:

- The target metrics is a calculation based on reach & engagement; therefore, shows that have low engagement will have lower ratings even if reach is held equal. There is no way to quantify how engaging (good/bad) a TV show is before it premieres
- I am working with a small data set, which is specifically for Viacom. It will likely not generalize well for Viacom specifically or outside of Viacom shows. In addition, it the small data set will limit the number of features I can include in the model.

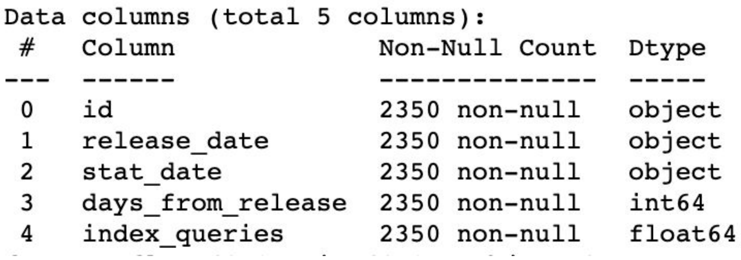
The Original Data

TV Ratings



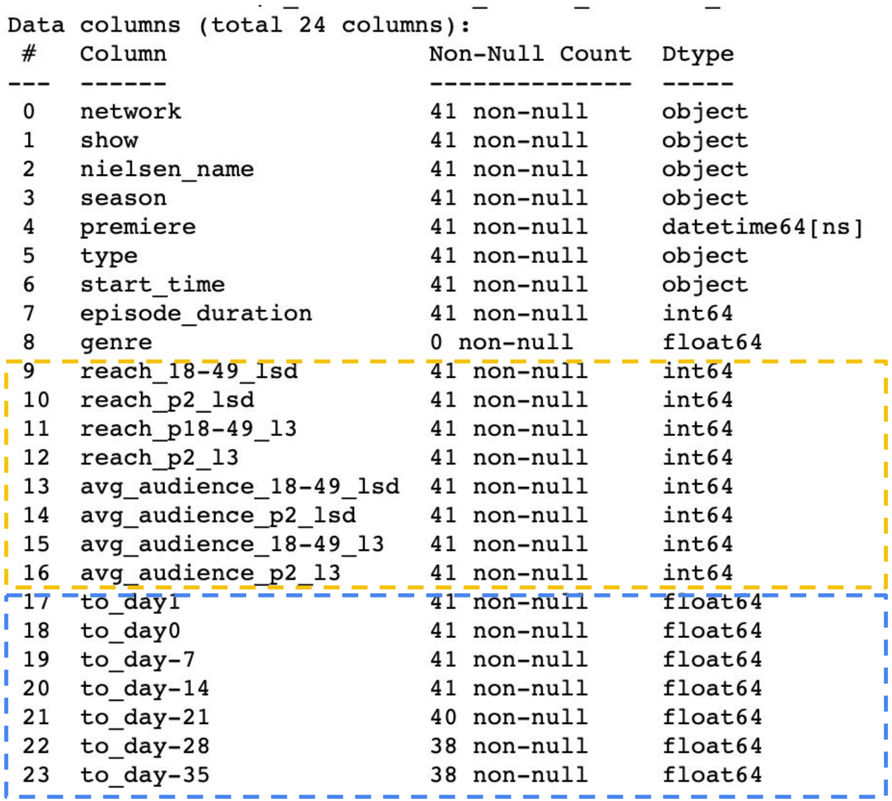
```
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   network                 41 non-null     object
 1   show                    41 non-null     object
 2   id                      41 non-null     object
 3   nielsen_name            41 non-null     object
 4   season                  41 non-null     object
 5   premiere                41 non-null     datetime64[ns]
 6   type                    41 non-null     object
 7   start_time              41 non-null     object
 8   episode_duration        41 non-null     int64
 9   genre                   0 non-null      float64
 10  reach_18-49_lsd         41 non-null     int64
 11  reach_p2_lsd            41 non-null     int64
 12  reach_p18-49_l3         41 non-null     int64
 13  reach_p2_l3             41 non-null     int64
 14  avg_audience_18-49_lsd  41 non-null     int64
 15  avg_audience_p2_lsd     41 non-null     int64
 16  avg_audience_18-49_l3   41 non-null     int64
 17  avg_audience_p2_l3      41 non-null     int64
```

Google Search



```
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 2350 non-null   object
 1   release_date       2350 non-null   object
 2   stat_date          2350 non-null   object
 3   days_from_release  2350 non-null   int64
 4   index_queries      2350 non-null   float64
```

The Combined Data



```
Data columns (total 24 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   network                41 non-null     object
 1   show                   41 non-null     object
 2   nielsen_name           41 non-null     object
 3   season                 41 non-null     object
 4   premiere               41 non-null     datetime64[ns]
 5   type                   41 non-null     object
 6   start_time             41 non-null     object
 7   episode_duration       41 non-null     int64
 8   genre                  0 non-null      float64
 9   reach_18-49_lsd        41 non-null     int64
 10  reach_p2_lsd           41 non-null     int64
 11  reach_p18-49_l3        41 non-null     int64
 12  reach_p2_l3            41 non-null     int64
 13  avg_audience_18-49_lsd 41 non-null     int64
 14  avg_audience_p2_lsd    41 non-null     int64
 15  avg_audience_18-49_l3  41 non-null     int64
 16  avg_audience_p2_l3     41 non-null     int64
 17  to_day1                41 non-null     float64
 18  to_day0                41 non-null     float64
 19  to_day-7               41 non-null     float64
 20  to_day-14              41 non-null     float64
 21  to_day-21              40 non-null     float64
 22  to_day-28              38 non-null     float64
 23  to_day-35              38 non-null     float64
```

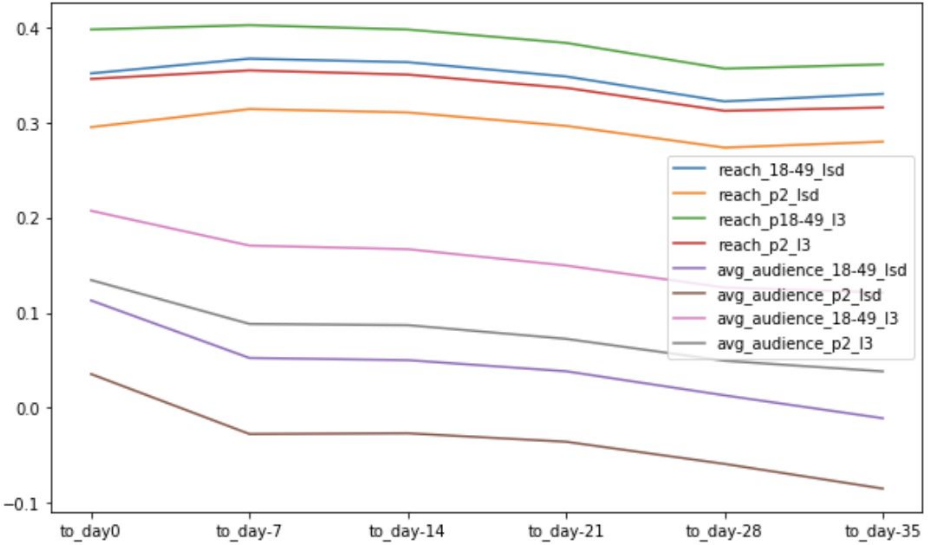
Two Core Questions:

Q1: What metrics has the strongest relationship & what should we try to predict?

Q2: How far out can we can confidently predict the result?

EDA

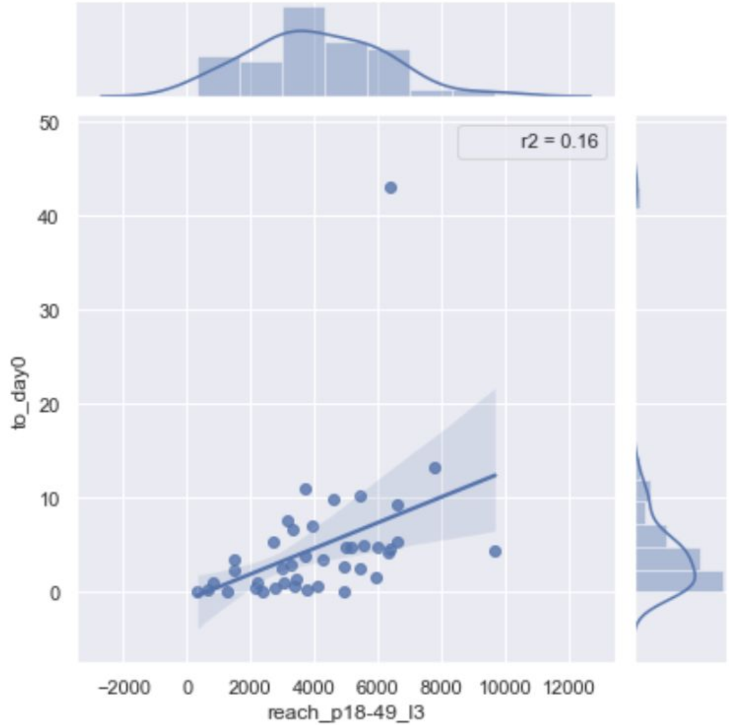
Q: What metrics has the strongest relationship?

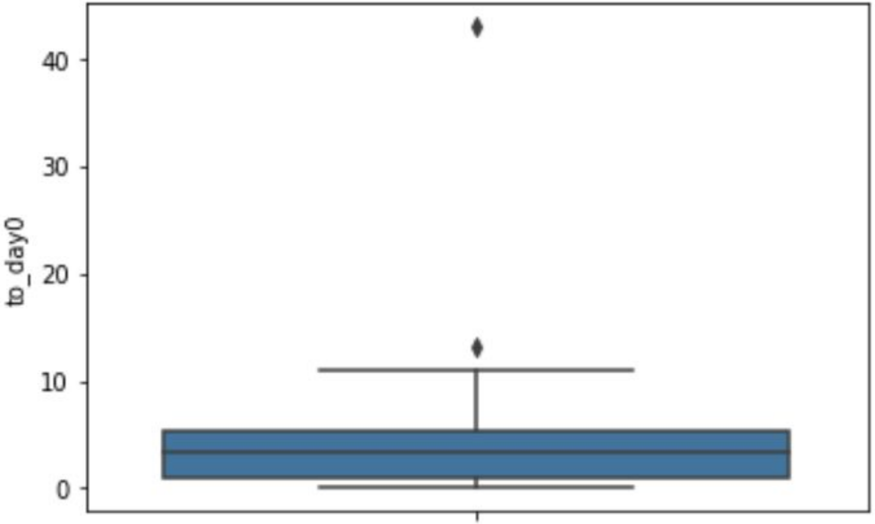


Findings

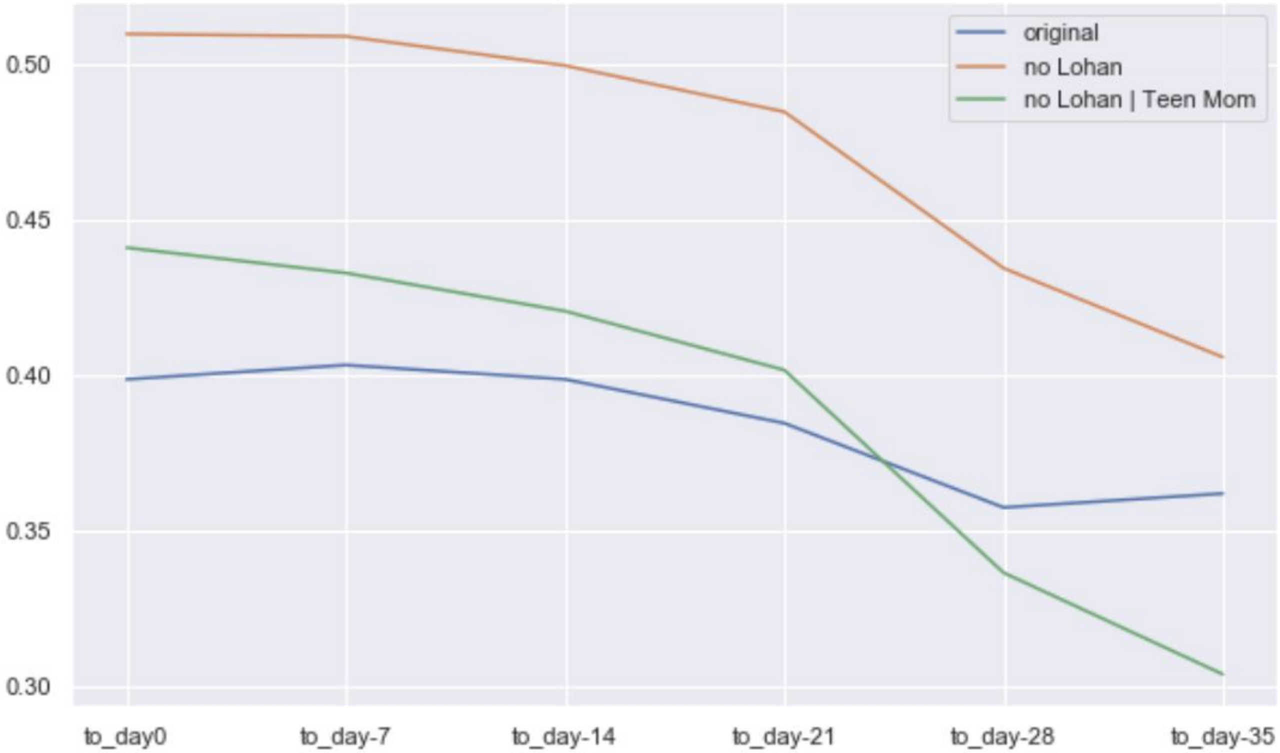
- Overall, Reach P18-49 L+3 has the strongest relationship between with Search
- All Reach related metrics showed a stronger correlation than any of the Avg. Audience metrics
- However, AA P18-49 L+3 had the strongest relationship for an Avg. Audience Metrics, indicating that Search Data is more likely to relate to the 18-49 audience than p2+
- the relationship between search and the metrics seems to remain relatively constant from -14 to 0, meaning that we may be able to make meaningful predictions two weeks out
- However, relationship is very weak, especially in comparison to similar analysis

There's outliers, are they skewing results?





A stronger relationship when we remove outliers



Are there any interesting Dimensions?

- Network
- Genre
- New/Returning/Tentpole
- Day of Week
- Time slot
- Seasonality

Modeling

What did I do?

What I Did

- OLS
- OLS w/ Cross Validation
- Lasso w/ Cross Validation
- Normalized Lasso w/ CV

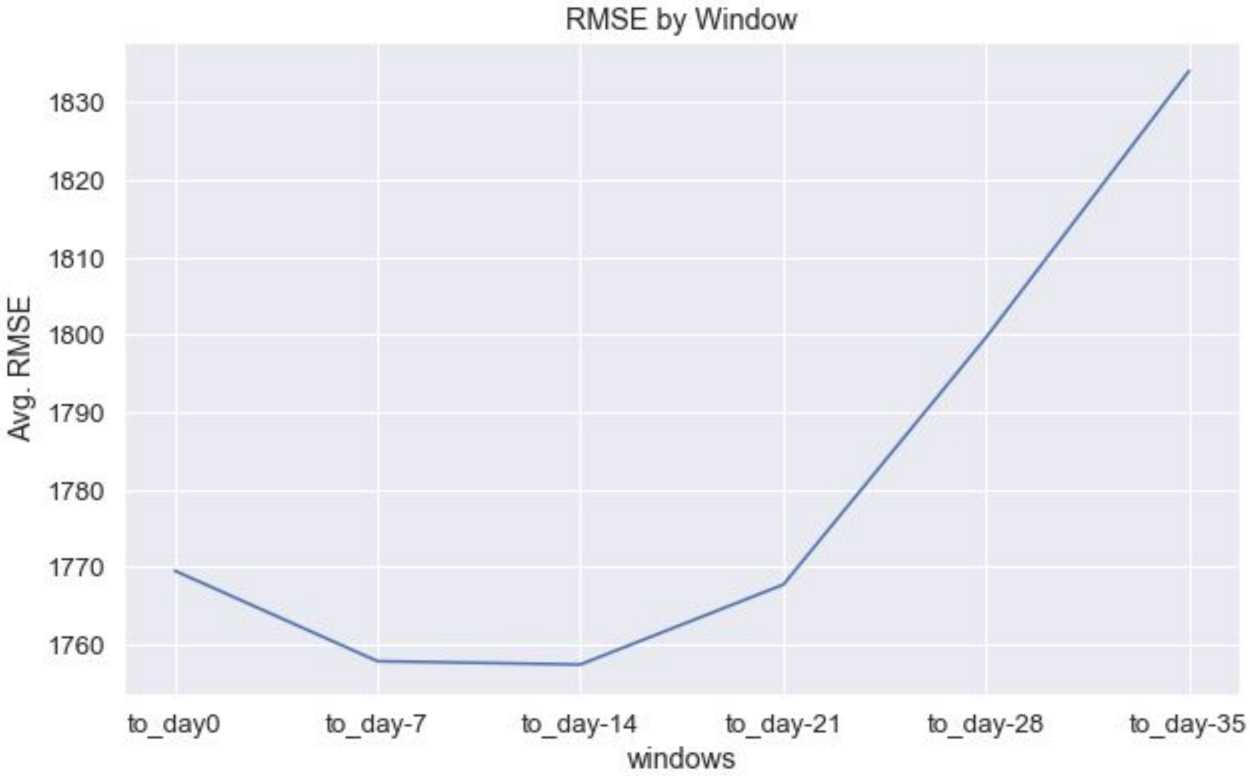
What I still want to do...

- Decision Tree Regressor
- Neural Networks

Results: What Model is most predictive?

Model	**RMSE**
OLS	1524.34
OLS CV (cv=5)	1740.98
Lasso	1740.95
Lasso CV (cv=5)	1782.66
Normalized Lasso CV (cv=5)	1782.21

Results: How far out can we predict



Thoughts going forward...

Conclusions

- Strong relationship between TV Ratings & Search Volume; can be used a strong indicator of future success
- However, data set was limiting: we could 1) not include a lot of features 2) Try more complex models like Decision Trees and Neural Networks, which resulted in low predictive power of mode: RMSE: 1.7M

Next Steps

- Next Steps: Expand data set outside of Viacom brands so we can include more features and test other models and improve predictive power