Search Engine Optimization Analysis

IST718 Big Data Analytics
Spring 2024
Ryan Richardson, Rodrick Blanton, Maya Davis

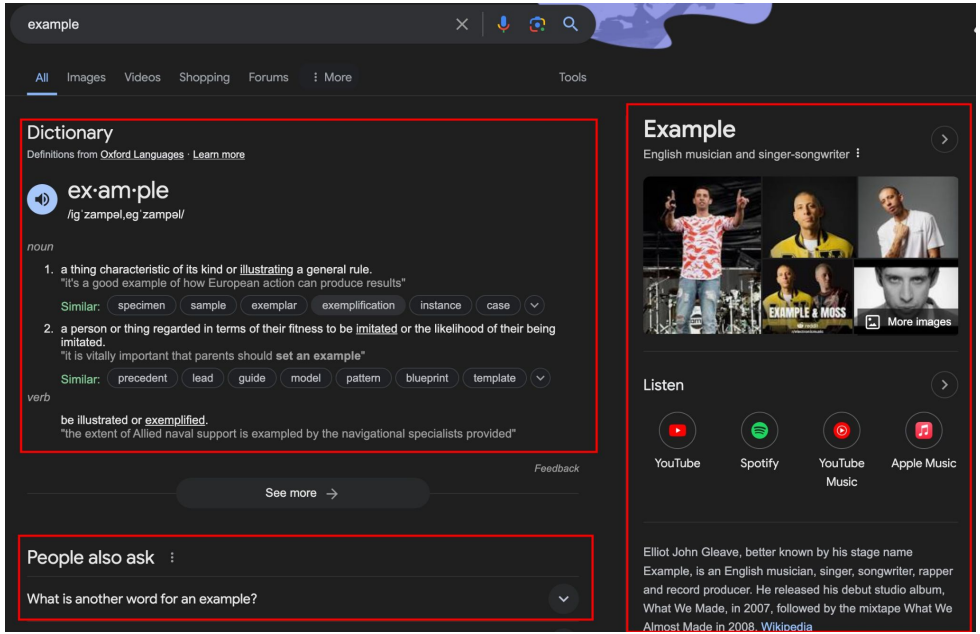
Introduction

- SEO market worth an estimated $68.27 billion[1]
- SEO costs for businesses can vary dramatically depending on the size of the business and its SEO needs. This can range anywhere from $500 a month for a small business to more than $10,000 a month for some enterprise operations.[2]
- Hypothesis: Costs can be minimized and ROI maximized through data science techniques
    - Approach:
        - Use FAISS to cluster queries and identify which clusters are the most valuable
        - Use machine learning models to predict which query SERP features should be prioritized for SEO strategy

1) https://www.emergenresearch.com/industry-report/search-engine-optimization-market#:~:text=The%20global%20Search%20Engine%20Optimization,factor%20driving%20market%20revenue%20growth.
2) https://www.webfx.com/seo/pricing/

Data Overview

- 150,000 rows with 5 columns
- “Keyword” - the search query itself
- “Volume” - how much US traffic navigates to company page from search query
- Global Volume - how much global traffic navigates to company page from the search query
- Traffic Potential - how many times the search query is run
- SERP features - list of additional features on the search query page including:
    - Bottom ads, Knowledge card, Video preview, Top stories, Videos, Thumbnail, Shopping results, Image pack, People also ask, Featured snippet, Paid sitelinks, Top ads, Sitelinks, Local pack, Knowledge panel, Tweets



Data Cleaning, Preparation, and Feature Engineering

- Data is proprietary and was shuffled to protect company assets
- Created some nonsensical rows where “Volume” exceeded “Traffic Potential.” Additionally some missing values.
    - Since the data was already shuffled and hypothetical, we opted for simply interpolating the mean into the “Volume” and “Traffic Potential.”
    - Would want a more robust approach in a production environment, especially since the data has many extremely high outliers
- Created a new “Opportunity” column as difference between “Volume” and “Traffic Potential” to identify particularly valuable queries and opportunities to refine SEO strategy
- SERP Features column is a single column and values are strings. Each features is separated by a comma. Split the column on the commas and one hot encoded each feature.
- The column also had missing values. In order to build a robust proof of concept, we opted to preserve the distribution of the SERP features column by sampling from the distribution and interpolating based on the sample.

Clustering

Approach: FAISS clustering with almost 200 predefined clusters

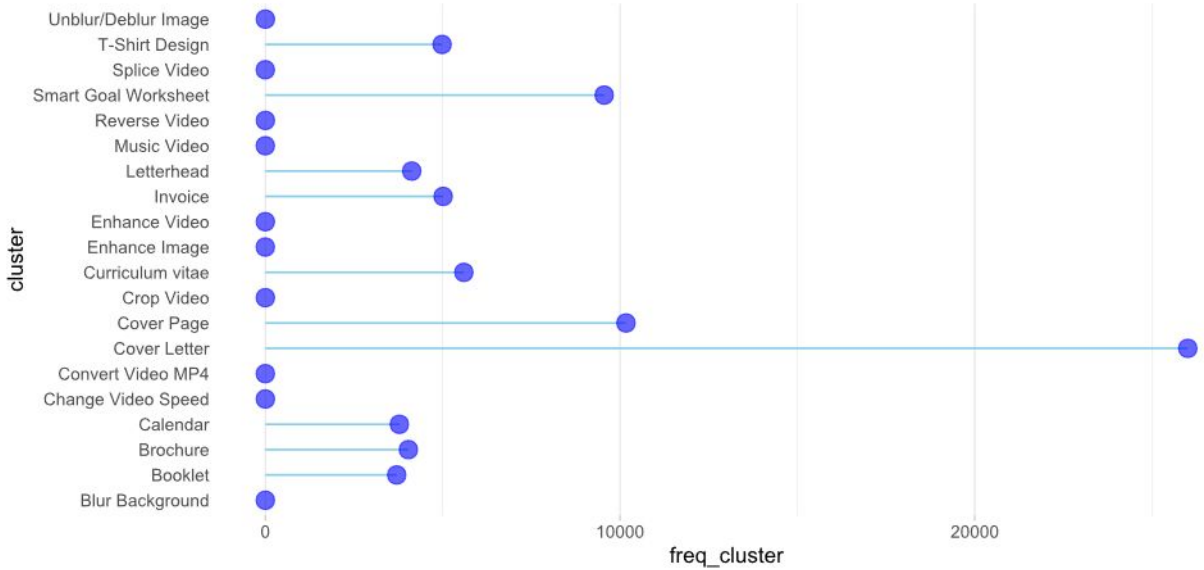
Reason behind using clustering:

1. Identifying high-value keywords to group similar queries to pinpoint keywords that are frequently searched together, indicating their relevance and potential value.
2. Understanding user intent to optimize content.

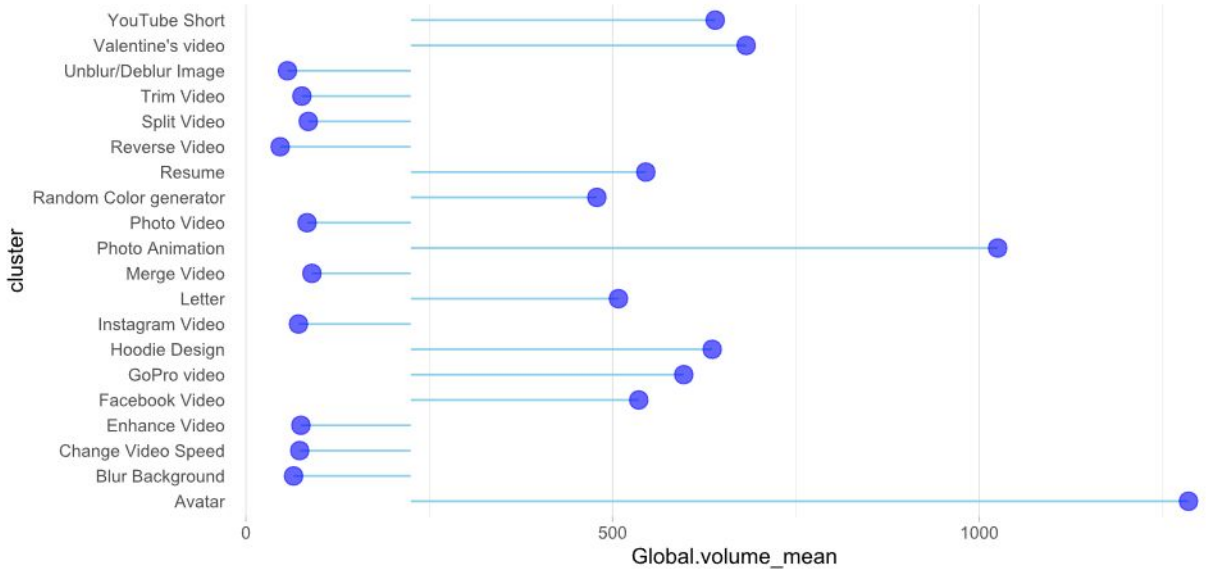
Clustering outcome

	cluster_name	queries	count	total_msv
0	Content Scheduler	home watch checklist template, timeboxing temp...	855	98790
1	Advertisement	social media advertisement template, blank adv...	118	9710
2	Banner	linkedin banner template 2017, etsy banner tem...	1453	157220
3	Flyer	flyer template free download, toastmasters fly...	1036	88940
4	Logo	nfl logo template, world series logo template,...	289	34550

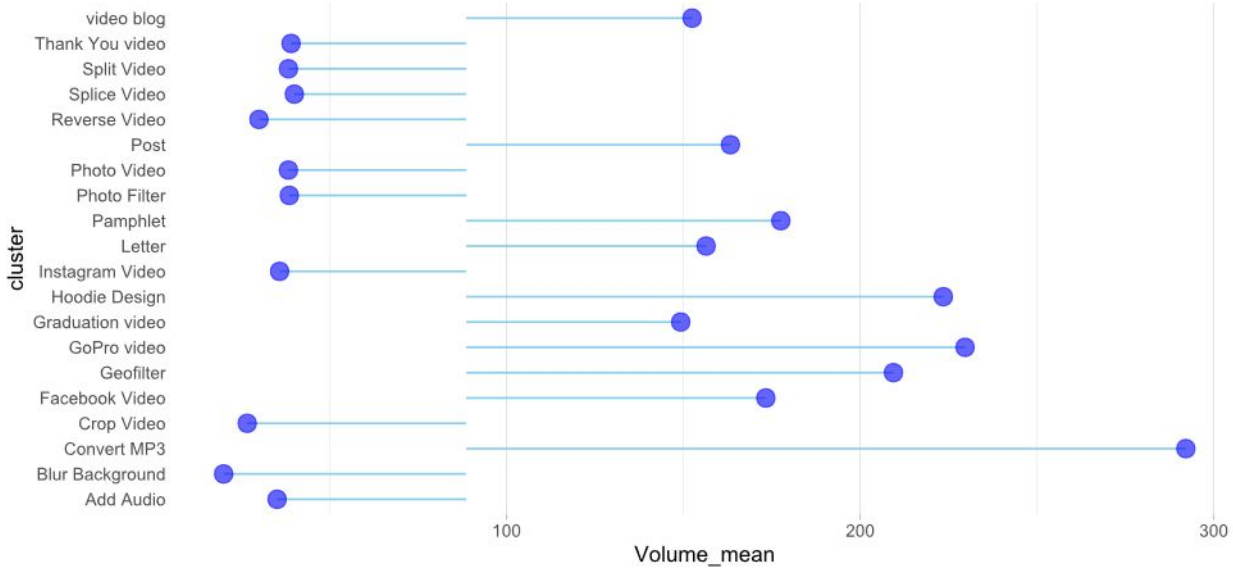
Top and Bottom 10 Clusters by Frequency



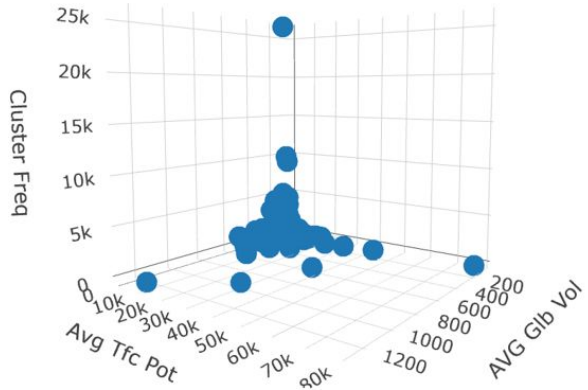
Best and Worst Performers by Global Volume

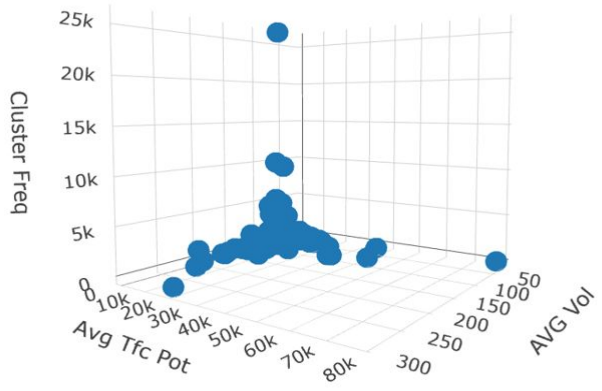


What is the impact locally?

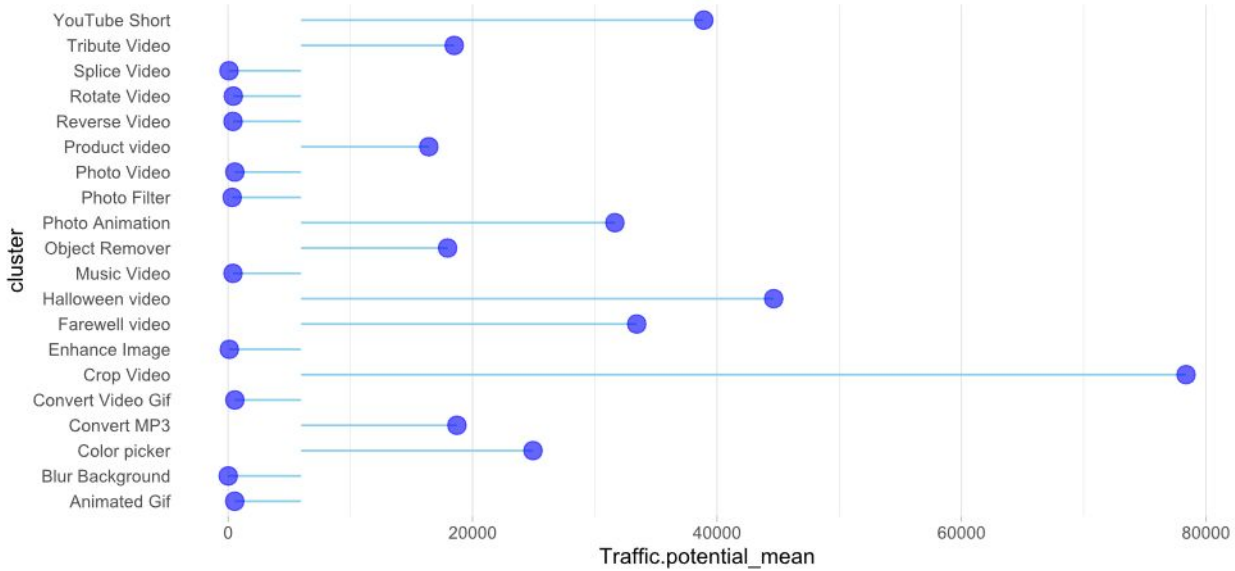


Measuring Terms

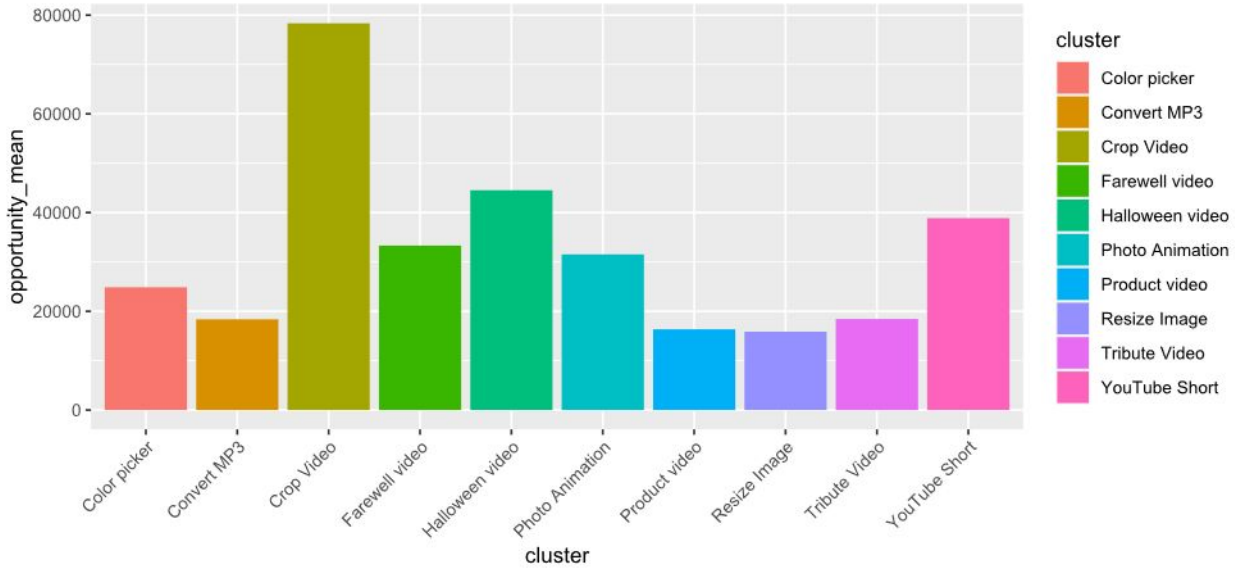




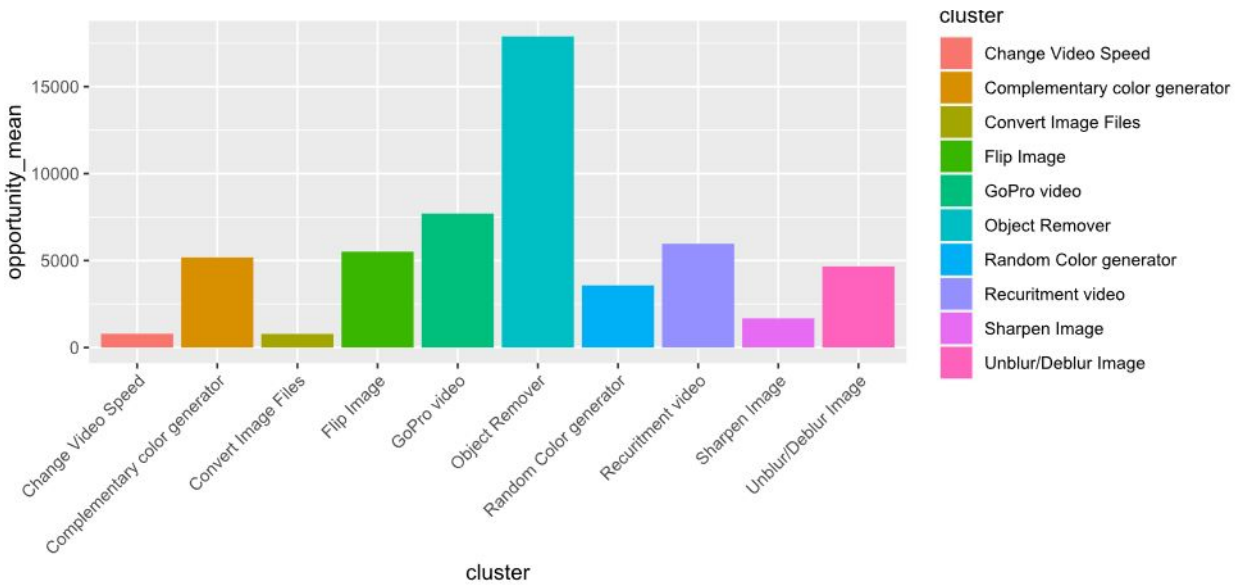
Top and Bottom 10 Clusters by Potential



Opportunity of the Best Assigned Clusters

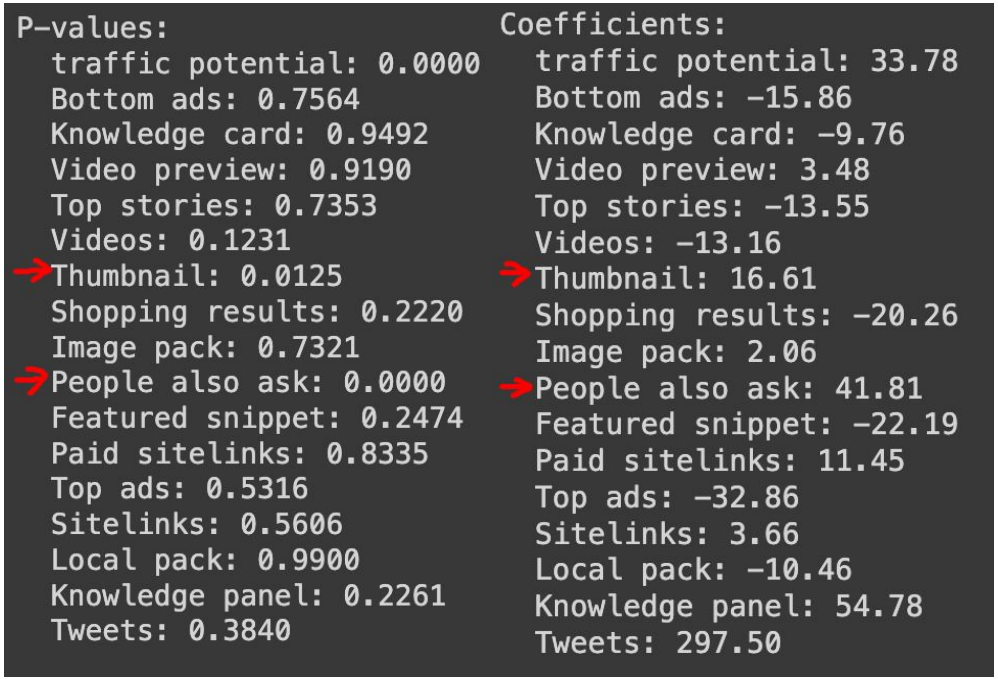


Opportunity Below the Line



Regression Modeling

- Created 3 regression models to predict “Volume”
- Created a baseline linear regression model with all SERP features included, along with traffic potential, to identify most important SERP features in predicting volume.
  - R-Squared: 0.0023
  - Mean Squared Error: 371,667
- Created second linear regression model with just those features included
  - R-Squared: 0.0022
  - Mean Squared Error: 371,668
- Created a Gradient Boosting Tree Regressor model with selected features
  - R-Squared: 0.0132
  - Mean Squared Error: 367,662



Classification Modeling

- Bucketized “Volume” into tiers and test models on classification accuracy
- Challenge: Identifying where to create tiers due to outliers. Began with 0-50, 51-100, 101-150, 151-200, 201+
- Built two classification models - baseline logistic regression model and a random forest classifier with selected features to predict Volume tiering
- Logistic regression model: 53.51% accuracy
- Random Forest Classifier: 58.84%

Logistic Regression:

```
Confusion Matrix:
[[2.4014e+04 3.0000e+00 0.0000e+00 0.0000e+00 1.6000e+01]
 [1.4784e+04 0.0000e+00 0.0000e+00 0.0000e+00 7.0000e+00]
 [1.6280e+03 0.0000e+00 0.0000e+00 0.0000e+00 2.0000e+00]
 [1.4210e+03 0.0000e+00 0.0000e+00 0.0000e+00 0.0000e+00]
 [3.0110e+03 0.0000e+00 0.0000e+00 0.0000e+00 1.0000e+01]]
```

Random Forest:

```
Confusion Matrix:
[[1.4886e+04 1.1230e+03 0.0000e+00 0.0000e+00 0.0000e+00]
 [6.9900e+03 2.8010e+03 0.0000e+00 0.0000e+00 0.0000e+00]
 [1.1010e+03 3.0000e+00 0.0000e+00 0.0000e+00 0.0000e+00]
 [9.5700e+02 2.0000e+00 0.0000e+00 0.0000e+00 0.0000e+00]
 [2.0880e+03 0.0000e+00 0.0000e+00 0.0000e+00 0.0000e+00]]
```

Conclusion and Next Steps

- Clustering unlocks larger cluster optimization vs keyword level
  - Saves time and resources
  - Ensures that the content is rich and comprehensive, which search engines favor
- Models offer early indications of value for further testing. While data has been shuffled and results should be taken with a grain of salt, we were able to identify two key SERP features
- Conversation with business stakeholders to identify whether regression or classification models are preferable, and if classification is preferable, how should tiering be designed
  - Both ensemble models are untuned, but still show definitive improvement over the simpler modeling methods. Once business stakeholders indicate preference, hyperparameters for the ensemble models will need to be tuned