# 

| Search Engine Optimization Analysis  IST718 Big Data Analytics  Spring 2024  Ryan Richardson, Rodrick Blanton, Maya Davis |
| --- |

In the digital age, search engine optimization (SEO) is a critical component for any company looking to enhance its online presence and attract more visitors to its website. The primary objective of SEO is to improve a website's ranking on search engine results pages (SERPs), which can significantly impact the visibility and traffic of the site. This paper explores the use of big data analytics to enhance SEO strategies, focusing specifically on clustering queries and analyzing SERP features.

One effective approach to improving search rankings is clustering similar queries together. Clustering involves grouping search queries that exhibit similar characteristics or intent. By analyzing these clusters, companies can identify common themes and patterns in user search behavior, which can inform content creation, keyword targeting, and overall SEO strategy. Clustering queries can help to do many things, but we’ll specifically look at 2:

1. Identifying high-value keywords to group similar queries allows companies to pinpoint keywords that are frequently searched together, indicating their relevance and potential value.

2. Understanding user intent to optimize content. Clusters can reveal the underlying intent behind search queries, enabling companies to tailor their content to better meet user needs. By understanding the commonalities among grouped queries, companies can create content that addresses a broader range of related topics, improving the likelihood of ranking for multiple keywords.

SERP features are any elements on a search engine results page that go beyond the traditional organic search results. These features include knowledge panels, local packs, featured snippets, images, videos, and more. Analyzing SERP features is crucial for several reasons:

1. Visibility to optimize their content to appear in these features, increasing visibility and click-through rates.

2. Competitive Analysis to gain insights into what competitors are doing and identify opportunities to outperform them.

3. Increase in traffic Potential. For example, featured snippets often capture a significant portion of clicks. Knowing which features are present can help prioritize SEO efforts.

In this project we’ll leverage big data analytics to cluster queries and analyze SERP features to provide valuable insights that help any company (in our case Adobe) refine their SEO strategies. By understanding user intent and optimizing for key SERP features, companies can enhance their online presence, improve search rankings, and achieve their digital marketing goals.

# Feature engineering

The data itself consists of roughly 150,000 search queries and data associated within them. Apart from the query itself, additional columns include the volume of clicks that route to our pages from each search query produced within the United States, the global volume produced by each search query, the total number of times that particular query is run (the traffic potential), and then a column with a list of SERP features associated with each query.

The data was proprietary and in order to protect business assets we began by shuffling the data. This produced the unfortunate side effect of producing some rows where there were nonsensical data combinations - where volume exceeded total traffic potential. In order to solve for this, we interpolated the mean of the volume and traffic potential columns into the respective values where this was the case. We also performed this interpolation on missing values. In a normal setting, a more robust methodology would be necessitated as this would warp the typical distribution of the data (although, it would probably have lower significance as we would only be interpolating on the relatively few missing values instead of all of the values where volume exceeded traffic potential as well). For our purposes, however, this was sufficient to build a proof-of-concept since the data had already been manipulated from its real-world setting.

The list of values in the SERP features column presented its own challenge to manipulate. We had several missing values and a list of features for each row (a single string, with the various features separated by a comma). We began by splitting the SERP features column on the comma and captured each unique feature in the column. This enabled us to one-hot encode the SERP features column creating numeric inputs for modeling. In order to avoid creating skew in the SERP features, we interpolated the missing values from the distribution in the column (essentially, sampling from the column and interpolating that into the missing values). In a production setting, it would probably be worth investigating further into why there are missing values in the data to ensure that this is the correct approach in handling these missing values.

Lastly, we also created a new “Opportunity” column in the dataset. This was created simply by taking the difference between Volume and Traffic potential as a way to easily demonstrate the missed value from the various queries and understand how to better optimize our SEO efforts.

# Clustering

To improve our SEO strategy, we conducted an extensive analysis of search keywords and clustered them into approximately 200 distinct groups. Each cluster represents a set of queries that share common themes or user intents. To achieve this, we utilized the FAISS (Facebook AI Similarity Search) module to efficiently encode the keywords and calculate the distances between them, facilitating the clustering process.

Our approach took a few tries in order to fine tune threshold to group keywords most logically. After encoding all the keywords (user queries ie search terms) and cluster names, we calculated the distance between keywords and clusters, and chose the smallest one. However, if the smallest distance was above our threshold 1.412 we moved those terms into the “uncategorized” group.

Choosing a threshold is very important to create clusters with queries that have common themes within to tailor our content to address specific user needs and queries. For instance, if a cluster reveals that users are frequently searching for information on a particular product feature, we can create detailed content that thoroughly covers this topic. If the threshold is too high, it will group unrelated queries together, if too low, it will keep a lot of them uncategorized.

# Visualization

As mentioned in the introduction the original dataset contains 150K rows of 5 features. The keywords come from search queries and different combinations of those words in queries. Volume and Global.volume represent actual traffic for those queries within the United States and around the world respectively while traffic potential holds the expectation for volume.

'data.frame': 150000 obs. of 5 variables:

$ Keyword: chr "resume template" "cover letter template" ...

$ Volume: int 90 50 150 40 20 30 20 20 30 20 ...

$ Global.volume: int 90 60 300 60 20 80 20 20 150 30 ...

$ Traffic.potential: int 2900 60 70 20000 NA 1400 NA 40 10 30 ...

$ SERP.Features: chr "People also ask,Image pack"...

Once the FAISS clustering has been implemented there are two new datasets for assigned and unassigned clusters. As mentioned earlier, the unassigned queries/keywords have a cluster but fall below the distance threshold.

'data.frame': 149899 obs. of 7 variables:

$ Keyword: chr "\"\"\"change-of-name\"\"...

$ Volume: int 30 30 30 60 60 20 30 20 30 40 ...

$ Global.volume: int 30 90 50 60 100 20 30 150 100 70 ...

$ Traffic.potential: int 150 5000 5200 40 70 NA 800 NA 100 20 ...

$ SERP.Features: chr "Image pack,Thumbnail,Sitelinks"

$ distance: num 1.28 1.03 1.15 1.48 1.29 ...

$ cluster: chr "Letterhead" "Cover Letter"...

To simplify the data description, it is necessary to summarize the numeric columns. The dataset is grouped by cluster and summarized. This gives a new structure of 193 rows representing the unique clusters. Those clusters can then be measured for frequency and added to the summary.

'data.frame': 193 obs. of 12 variables:

$ cluster : chr "Add Audio" "Add Caption" "Add Text on Photos" "Advertisement" ...

$ Volume\_mean : num 35.1 50.5 104.4 85.2 81.8 ...

$ Volume\_sd : num 21.3 72.4 222 193 212.2 ...

$ Global.volume\_mean : num 305 126 184 191 181 ...

$ Global.volume\_sd : num 1463 260 389 455 564 ...

$ Traffic.potential\_mean: num 5634 8450 8719 6137 6191 ...

$ Traffic.potential\_sd : num 18868 20471 24449 18806 46644 ...

$ distance\_mean : num 1.31 1.29 1.24 1.15 1.25 ...

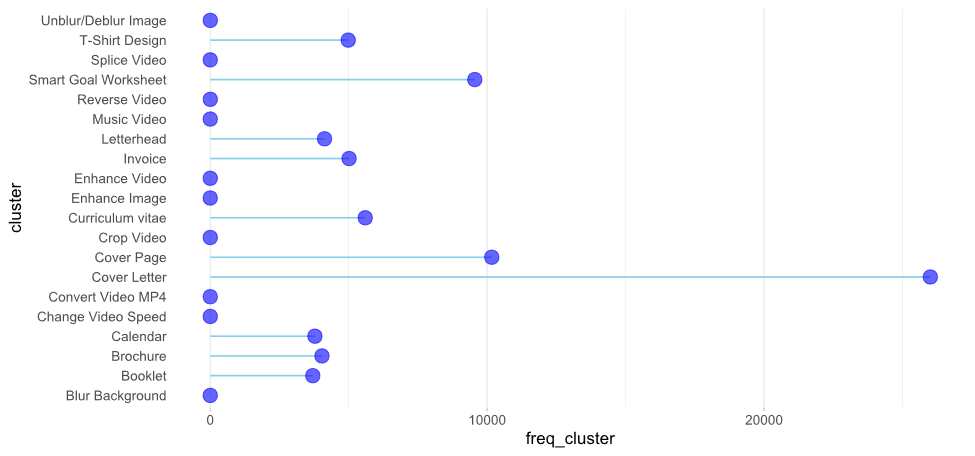
$ distance\_sd : num 0.1367 0.2572 0.0967 0.152 0.1847 ...

$ opportunity\_mean : num 5596 8394 8606 6044 6102 ...

$ opportunity\_sd : num 18873 20480 24438 18758 46637 ...

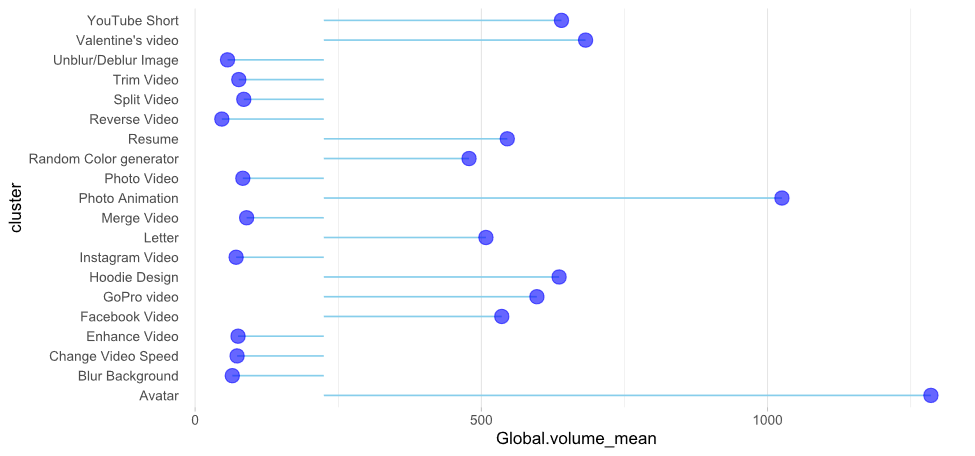
$ freq\_cluster : int 39 20 41 124 797 485 9 40 107 37 ...

Top and Bottom Clusters by Frequency

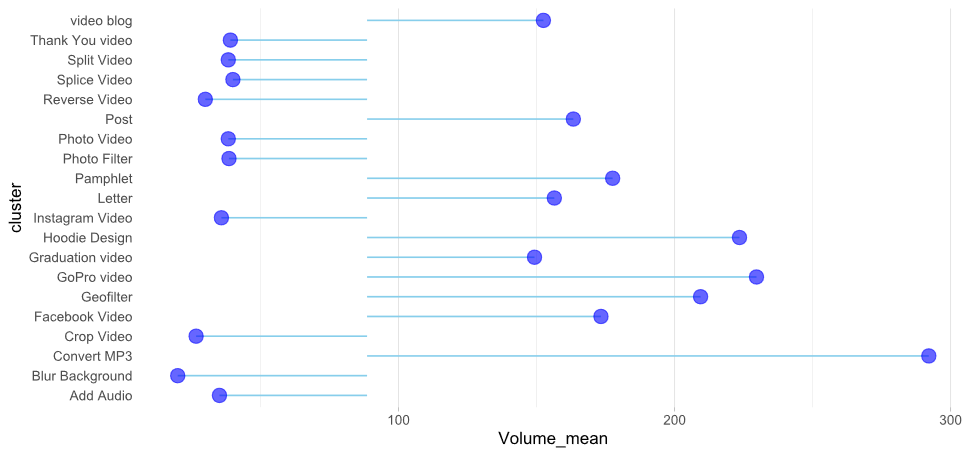


Top and bottom clusters by frequency is a measure of how many queries are represented by a cluster. The above graph represents the top 10 highest groups and the bottom 10 with a median at about 150. Due to the scale the median appears to be at 0 but it’s not. All of the lowest clusters technically fall below the median.

Top and Bottom Clusters by Mean Global Volume

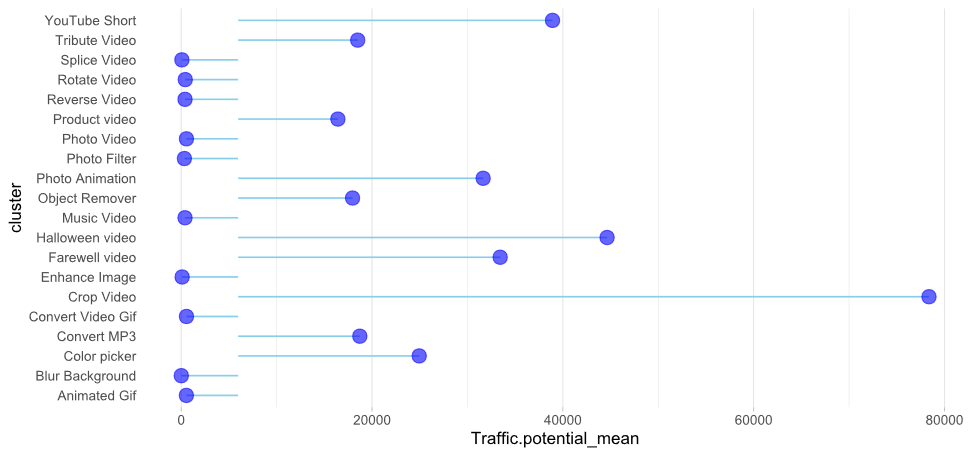


Another measure to look at is the average global volume by cluster. Here the median is roughly 250 hits. Some of the top outliers are Avatar and Photo Animation. Extending towards zero are bottom global performers like Blur Background, Reverse Video, and Unblur/Deblur Image. There’s likely no opportunity in those and they can be ignored while the higher outliers are worth further investigation.

Top and Bottom Clusters by Mean Local Volume

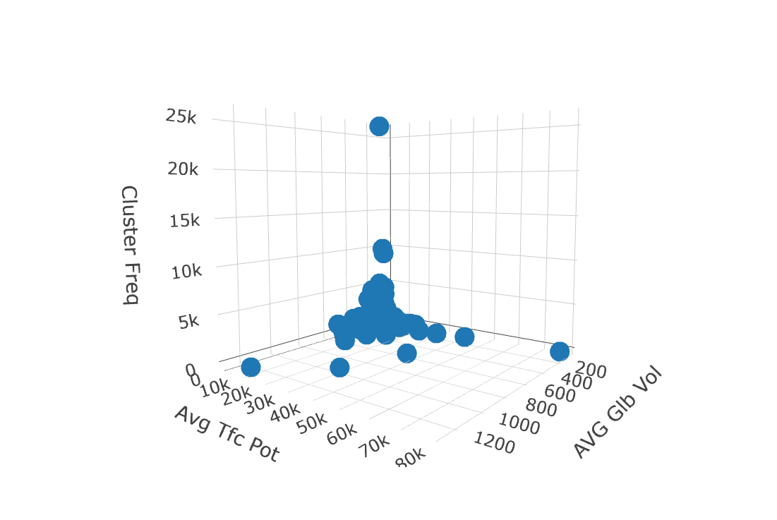
However, what’s more important is the local volume. This is volume within the United States and matches the scale of Traffic volume, which is how opportunity may be derived. Notice that the outliers look very different from global volume and crop video, which wasn’t an outlier globally, is an outlier within the US. Note also that blur background seems to consistently round out the bottom further indicating this isn’t a cluster worth investing resources.

Top and Bottom Clusters by Mean Traffic Potential

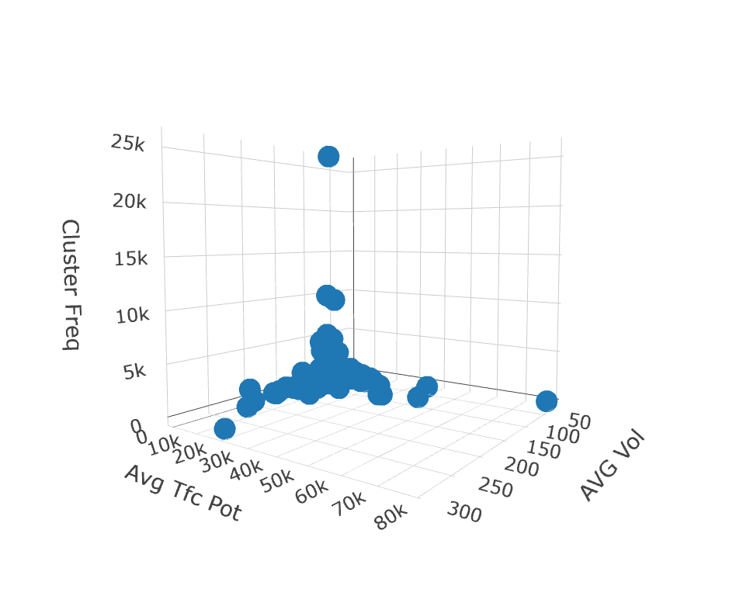


Looking at traffic potential, our eyes are immediately directed back to Crop Video. One may directly surmise the immense opportunity with this cluster. Local volume is below 300 while the traffic potential is just below 80,000. All the previous graphs have shown the top and bottom 10 clusters. Therefore, of all the clusters, crop video likely has the greatest opportunity and it would be worth investing heavily in developing targeted content.

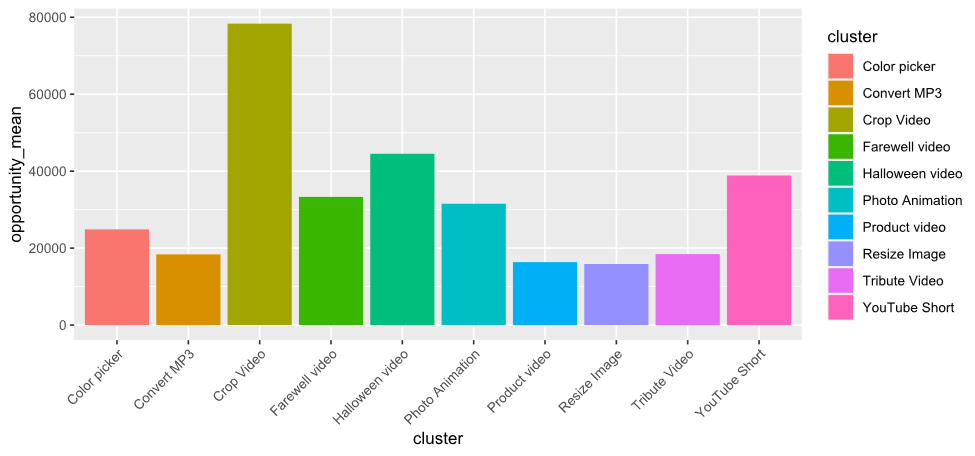
Relationship of Frequency Average Potential Traffic and Global Volume



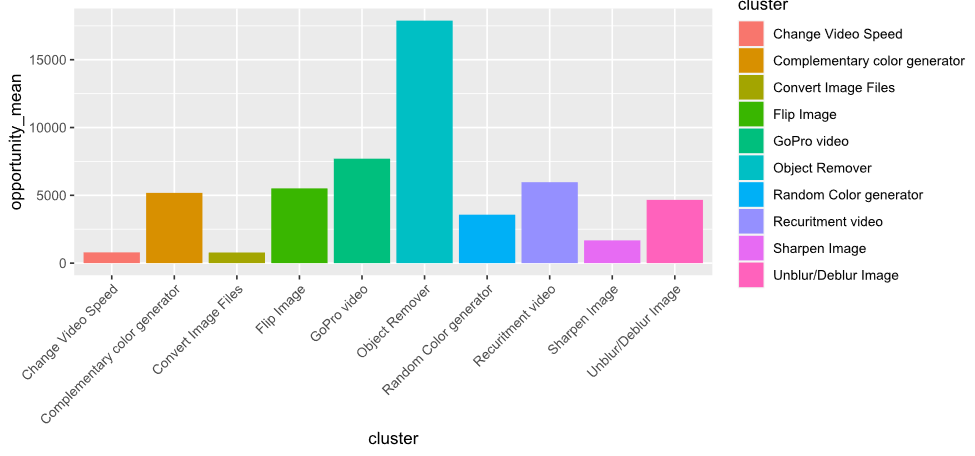
One thing to note is the relationship between three of the variables. It’s possible that cluster frequency, traffic potential, and volume (be it global or local) are all too closely tied together; the meaning of the terms overlap. In both the graph above and the one below, if the cluster frequency is low, it’s likely the traffic potential and volume will be low also. That’s a logical statement because the terms are so close in meaning.

Relationship of Frequency Average Potential Traffic and Average Local Volume

(*see paragraph above*)



As expected, the best or most opportunity is developing content for video cropping. Halloween videos and YouTube shorts also show some promising opportunities. However, something interesting develops when looking at the difference between assigned clusters and unassigned clusters. Remember that unassigned clusters are simply those that fell below the determined threshold. Note that the bottom of the top 10 assigned clusters have opportunities between roughly 15-18,000.



Now consider the unassigned clusters, those that fell below the threshold, and note that object remover also has an opportunity between 15-18,000. What does this mean? Another measure to evaluate is the distance. How far from the cluster terms are the keywords and SERP features? We established a threshold of 1.412. The scale extends from 1 to 1.8. If a cluster falls close enough to the threshold and displays high enough opportunity should that raise some questions. Two questions were, in fact, brought forward.

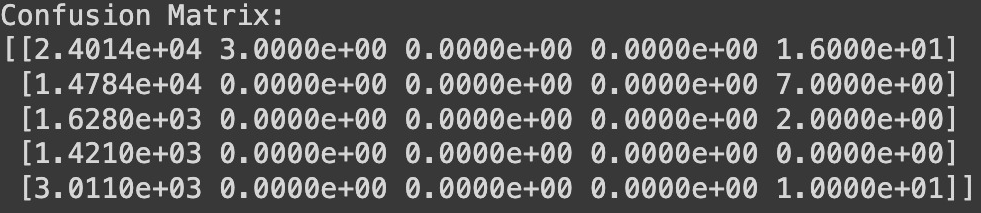
Object remover has an opportunity matching the lower end of the top 10 best opportunity assigned clusters. 1) Should money be invested in developing content for object remover or figuring out FAISS matching for that category. Maybe some keywords belong in other clusters and, with further investigation, the opportunity would shrink. 2) Is the threshold too high? The distance for object remover is 1.44, which is relatively close given the scale. So, changing the threshold would modify any analysis provided to a client.

# Modeling

As a reminder, the data was shuffled, so the results should be taken with a grain of salt, but represent a potential proof-of-concept that seem to offer some cautiously optimistic indicators that could provide value in a standard setting. We ultimately built five predictive models from the data - three regression models and two classification models attempt to predict volume.

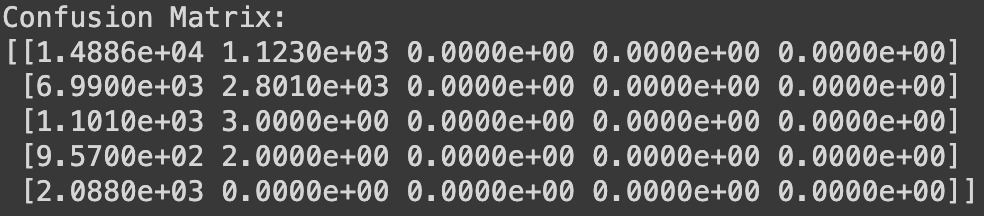
Beginning with our regression models, we initially started with a baseline linear regression model that took every SERP feature along with traffic potential as an input in order to try to predict volume. This model had a low R-Squared value (0.0023) and a mean squared error of 371,667. This indicates that the inputs of the model did not appear to account for the shift in volume. However some of the p-values for the inputs demonstrated significance - notably traffic volume, as well as the “People also ask” and “Thumbnail” SERP features, all with p-values less than 0.015. This provided an opportunity to create a second, refined linear regression model with just those features. This yielded a significantly smaller and simpler model with a negligibly worse performance (an R-squared value of 0.0022 and a mean squared error of 371,668). Given the similarity in model performance, we focused our remaining modeling efforts solely on these particular features. Our last regression model consisted of a Gradient-Boosted Tree Regressor. This model significantly improved the R-Squared value to 0.0132, however still yielded a high mean squared error of 367,662. Given the improvements in this last model, I would suggest at least running a test of the model on unshuffled data to see if it produced any stronger predictive power.

For the classification models, we bucketed the volume column into tiers. This process was challenging and admittedly, a little arbitrary. The volume column consists of mostly low values (values between 50 and 150 or so), but with quite a few extreme outliers reaching over 10,000, with several reaching even 50,000, 100,000 and even one over 200,000. Given the low range of the bulk of the data, we created tiers of 0-50, 50-100, 100-150, 150-200, and those over 200 (so 5 classes in total). We created a baseline logistic regression model that took traffic potential, People also Ask, and Thumbnail as inputs to try to predict which volume tier it would fall in. This produced a model with a 53.51% accuracy - not ideal, but significantly better than random guessing. Looking at the confusion matrix:



We see that the model is too sensitive to predicting either the 0-50 class or the over 200 class. This further demonstrates the challenges associated with the data distribution.

However, all the same we also tried to see if we could improve the model performance through a random forest classifier, and found initially encouraging results. An untuned random forest classifier without any hyperparameter tuning produced a model with a 58.84% accuracy. When examining the confusion matrix, we also see some interesting differences:



Whereas the first model favored predicting the low class or the high class, the random forest model predicts entirely within 0-50 or 50-100. While this again illustrates the challenges associated with the data distribution, the different preferences of each model, that also yield relatively high accuracy compared to the number of classes suggest there might be opportunity to improve this approach. Next steps should include conversations with business stakeholders to better understand volume tiering with communication around the challenges in the distribution of the volume to create a tiering system that best align with business strategy and goals as well as further hyperparameter tuning on the random forest model.

# Conclusion

With clustered queries, we streamlined our content optimization efforts. Instead of optimizing for individual keywords, we focused on optimizing for entire clusters and SERP Features. This means we now have an opportunity to create comprehensive content pieces that cover a range of related queries, reducing the need for multiple pieces of content. This not only saves time and resources but also ensures that our content is rich and comprehensive, which search engines favor.