# Zepto Assignment

~Abhigyan Tripathi(22085005)

**Problem Statement:**

The problem statement involved a dataset containing the car of products for specific queries and given cities ( in term of CITY\_ID). Our task was to build a ranking system based on given search term(query) and city.

**Approach:**

1. The basic idea of building a **ranking system** is to choose a relevance criteria or a **relevance score** and based upon that score e rank the datapoints
2. Using this basic idea of ranking , I first created a separate dataset containing all the numeric features and using that dataset found out the **correlation** of the features with the **is\_clicked** feature, this gave us a good hold on what features to use and what not.
3. Our next task was to create our own relevance score feature and this was done using the help of  **Logistic Regression**
4. Using Logistic Regression and **is\_clicked** as our target variable we found out the **Coefficients** belonging to the **top 7** features in the new data frame (selected using correlation coefficients calculated earlier)
5. Then using the coefficients we calculated the value of relevance score for each data point in the dataset
6. Now simply **sorting** the dataset based on **search\_term, city\_id** and **relevance score**, we get the desired rankings for a given search item and city id.
7. **Evaluation :** Evaluation could have been done using the **NDCG**(Normalised Discounted Cumulative Gain) but it was out of the scope of my system so I wan’t able to implement it.

**\*\* Initial Evaluation approach and why it failed on my system**

The initial approach for ranking was to use a **List-wise Ranking Model** which would use **Random Forest** to evaluate the Ranking by comparing the pair-wise relevance scores of the Products for a given search\_term and city and then rank them. Then the final rankings would have been evaluated using NDCG which works best for **List-wise** Ranking models.

It failed because the number of unique search\_terms (len(search\_term.unqiue)) was close to 2000 and there were 8 unique city ids and for each search term and city there were roughly 50 products to rank this led to the overall iterations in the loop exceeding 10^7 which caused failure in my case.

**Solution:** Using a GPU might work well in our case or using a POINT\_WISE ranking model.

\*\***Important Note:**

Even though we could have used **Embeddings** to describe the relationship between a given search query and product displayed, I did not perform it because a feature called **Query\_Product similarity** was already present in the dataset telling us about the **Cosine Similarity**  between the query and product , also it had a high correlation with the target variable and hence I thought that using Embeddings would make the model computationally expensive.

**##** Link to the Code (Code is present on my **Git-Hub** profile under the Zepto\_assignemnt repository):

**https://github.com/Abhigyan-Tripathi/Zepto\_assignment**