**Data Scientist Assignment Summary**

1. Problem Statement:

To provide score/rank to users based on browsing history and conversion data on e-commerce website to decide bidding logic for individual user.

1. Approach Taken:

Approach to solve this problem has been divided into below steps:

1. Data preparation

For creating training data was most challenging part for me in this exercise, to create training dataset I have followed below assumptions:

* At user level in a day minimum and maximum timestamp is duration that user spent on website/store
* Number of records associated with each user by website section are considered as number of times user has visited that page
* If user is purchasing any product, then he/she would be present in website browsing dataset

Along with this, to create target variable(conversion\_flag) as binary whether user has conversed/purchased product or not through website. For same, I have used final purchased dataset.

By using this dataset, how to get score/rank for each user to get purchase product? I have considered this problem as binary classification problem and developed conversion probability to get score for each. This probability tells us how likely it is for user to get purchase product from website. Higher the score better it is. This can be used in to decide biding logic for individual user

1. Feature generation

Based on above mentioned assumptions, I have generated the features for each user. Initially, I have tried using features from final purchased datasets, but those features were totally biased due it availability of them for conversed user.

1. Outlier detection

By using few charts/plots, I have eliminated few records from the dataset to improve performance of the models

1. ML modelling and final model selection

I have tried different classification models available in sk-learn. Results of these can be seen the ROC curve plot. Due to class imbalance in the dataset accuracy would not be good measure to select the final model or tune hyper parameters. So, selected f1\_score, AUC and ROC curve to evaluate different models and compare amongst them which are good evaluation metrics for imbalanced dataset.

1. Feature importance

Once final model is selected, by using same created the feature importance metrics for each of the training features.

1. Scope on improvement:

Nothing is perfect at first time, to improve this model there are multiple approaches or data preprocessing can be done. Please find below few point to do improvement in model:

1. If we get the product description of the product page which user is visiting that can be used to do text mining to identify more features which might help to improve accuracy.
2. As you look at the conversion rate of user it’s ~8% on one single day which has introduced the class imbalanced in the target variable. This can be tackled by introducing synthetic data points for minority class or based on some clustering techniques
3. I haven’t worked on normalizing variables in this exercise due to time constraint. That would be definite area of improvement
4. Evaluation of this implemented or productionized solution can be done by doing A/B testing
5. Interpretation of Results:
6. Users visiting to website section like cart, default, product and home, they have very high probability to purchase the product.