Intelligent Data Analytics (IDA-5103-001)

Reflection Paper by Power Rangers

The aim of this project was to predict the probabilities of the advertisement being clicked using various classification models. The data used for the project was obtained from the Kaggle competition Outbrain Click Prediction. Cross Industry Standard Process for Data Mining (CRISP-DM) was used to complete the project. It has 6 stages. Project Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.

Project Understanding: To predict whether the advertisement will be clicked or not.

Data Understanding: There were in total 20 features selected for final dataset used in modelling. There were several assumptions made as it was not specifically mentioned by Outbrain, such as the document id from Events described the current document or the source document and the document id in the Promoted content described the landing document or the document on which you will be lead to view once the ad is clicked on the current document. Such data understanding is necessary for any project. There needs to be a field expert in every project to provide guidance in the matters of data understanding.

Data Preparation: The data has 80 million rows in train file and 20 million in events page. The views file was 100 GB uncompressed. It was impossible to merge the dataset with members computer specs. Hence another machine with 32 GB RAM was borrowed where the data was merged into one file. But running the models with complete data was still impossible hence top 100000 rows were selected and 5000 rows were randomly chosen with 4000 of them having 60 % to 40 % ratio of no to yes in train dataset and remaining 1000 in test dataset. It was important to have at least 40 % of yes data to train the model to predict yes in the test dataset. Features used for modelling were selected gradually with more data exploration and model results and evaluation. The process was iterative. Dummification or on hot encoding was used to transform the continuous numerical data to categorical data in binary form. Difficulties were faced in dummification but achieved with the help of guidance from Dr. Nicholson. Feature selection was done based on outlier analysis, missingness of data, correlation between features. Feature engineering was performed on geolocation to separate the data into three separate features of country state and dma.

Modelling: Models were chosen based on knowledge of the model and the timeframe we had. It is always preferable to run logistic regression and random forest as base models to understand the output. Initial modelling results with 3 features were poor than a random model with AUC less than 0.5. It was then realized that the numeric data was considered as continuous numerical data while on the contrary it was a categorical data. The data was then transformed into binary categorical form. The results then improved with results greater than 0.5 in the range of 0.51 to 0.55. With increase in features (feature selection and feature engineering) the results improved to 0.66 of AUC for random forest. Some of the models like logistic regression and some cross validation and hyperparameter tuning were carried out but results were not obtained due to limitation on time and memory in the computer. Random forest, glmboost, gbm, black boost were the models applied for the project. Factorization machine s and B

Evaluation: AUC in ROC curve was used for model evaluation. This method does not require the need for cutoff selection for binary classification to evaluate the model. For a random model the AUC is equal to 0.5. Hence model can be assumed successful if the AUC is greater than 0.5.

Deployment: With the best model result as 0.66 we can say that modelling was successful and it can be applied for displaying the ads on document which have highest probability of being clicked. It is better than displaying random ads and being hopeful of it being clicked. With an increase in memory space (probably use a super computer or cloud computing) and more time, the results can be improved with inclusion of more data for training the model.

Feedback Summary: The feedback was given as comments in the PFD sent to Tatsumakifriends for peer review. They were satisfied with the amount of work done in the initial phase. They knew that the data was complex and hence were supportive of our work with less critic of the work. They liked the idea of using clustering for data exploration. They were hopeful of us applying different models which we mentioned in the plan forward section.

Feedback: Feedback received from the Tatsumakifriends was encouraging. They approved of our understanding of the data and the methods we used for data preparation and modelling. Though we had difficulties in applying one hot encoding, feedback from Dr. Nicholson helped us in applying dummification and understand the data better. They approved of our model selection and were excited to see the results, but due to time limitation we stick to the known models from the assignments and did not divulge in other models like factorization machines and Naïve Bayes.

Iterations: Since there were no critical and specific suggestions from the reviewers, iterations were self-deployed and are explained in detail in Data Preparation and Modelling section of the CRISP-DM.