Data Analysis

Presenting in this report is an analysis encompassing exploratory data analysis (EDA), data pre-processing, feature selection and engineering, model selection, model evaluation metrics, model tuning, performance monitoring, and the maintenance plan of a machine learning model meticulously designed, constructed, and deployed.

This model aims to predict customer inclination towards the "Add to Cart" action, accurately determining whether this action will occur based on the chosen or developed features from the provided dataset. Integral to achieving the Q1 business objective of enhancing the return on ad-spending, this model is crucial for the company utilizing Google Analytics. Utilizing the sample dataset from Google Analytics 360 on BigQuery, a model was crafted and implemented in VertexAI to forecast instances where a 'hit' would result in an event action of "Add To Cart." Initially, a training period of one month in March 2017 was selected, with the intention of seamlessly substituting the 12-month period (full dataset) post-validation. Given the dataset's magnitude, data cleansing and feature selection were employed to efficiently test and train the model. All such decisions have been meticulously documented in the corresponding Jupyter code notebook or within this comprehensive report.

Number of Rows – 1,641,689

Features – 65

Target – “Add to cart” (Action_type)

Data Preprocessing

Data with null values, duplicates was removed and we were left with 42 features.

Numerical, Boolean and categorical features were separated. “onehotencoder” was applied to the categorical features.

Feature Selection

- 10 features such as “hitnumber”, “visitID” etc. were selected based on the schema to be the unique features.
- Second set was made from the data table based on the schema.

Model Results

Model 1 results

1. Hypertuning

   Number of trials – 8
   Max number of parallel trials
   ROC-AUC curve
   Learn Rate – 0.01, 0.1
   Tree Depth – 5,6

2. Results

   Precision – 0.0
   Recall – 0.0
   Accuracy – 0.97
   F1 score – 0.0
   Log_loss – 0.14
   ROC-AUC – 0.74

Model 2 results

1. Hypertuning

   Number of trials – 8
   Max number of parallel trials
   ROC-AUC curve
   Learn Rate – 0.01, 0.1
   Tree Depth – 5,6

   Test rows - 116,444
   Train – 925,161
   Eval – 115,777

2. Results

   Precision – 0.017
   Recall – 0.89
   Accuracy – 0.62
   F1 score – 0.03
   Log_loss – 0.53
   ROC-AUC – 0.82

Interpretation

It was found out that the hits, pageviews contributed maximum to the model. Model 2 performed better than the Model 1 but it is still not the best due to the rush in creating these models. No other models could have been tried.