Analysis Report

1. Introduction

This analysis report presents exploratory data analysis (EDA), data pre-processing, feature selection and engineering, model choice, model evaluation metrics, model tuning, performance monitoring and maintenance plan of machine learning model designed, built, and deployed to predict customer propensity to perform the "Add to Cart" action. The model is to be used to accurately predict whether this action will be performed or not based on the features selected or developed from the given dataset. This model is essential as part of the Q2 business goal to increase the return on ad-spending of the company employing Google Analytics.

Using the google analytics 360 sample dataset on BigQuery, a model was developed and deployed in VertexAI to predict when a ‘hit’ will have an event action of “Add To Cart”. To start with, a 1-month period of July 2017 was chosen to train the model, with the goal of being able to easily substitute in the 12-month period (full dataset) after validating the model. Because of the dataset size, data cleaning and feature selection was used in order to test and train the model in a timely manner. All such decisions were documented in the corresponding jupyter code notebook or within this report.

2. Data Overview

- Total samples: 324,096
- Features: 311
- Target: 1

3. Data Cleaning

Missing values: Features containing only null values (172) were droped from analysis. Similarly, features containing "not available in demo dataset" (17) were droped from analysis. Features (7) clearly occuring after Add to Cart action were droped. Features related to page load time (11) were judged out of scope and were droped.

Remaining features (105), including those containing high null values were allowed to remain candidates for further consideration. This was compared against a new list of (58) candidates selected solely on schema and dataset documentation, resulting in a final refined list of (34) final feature candidates for exploratory data analysis.

Data transformations: Transormations were applied to all features as appropriate during the exploratory data analysis, model training, and model deployment.

Categorical encoding: One-hot encoding was applied to all for EDA of final (34) feature candidates. All categorical data was encoded as part of the BigQuery ML model training process.

4. Exploratory Data Analysis (EDA)

One-hot encoded final (34) candidate features derived from (311) initial candidates. Resulting (84) one-hot encoded features measured for linear correlation and mutual information (MI). From this, the (10) features ranking with the hightes MI scores were fed into the first model.

Visualized linear correlation of feature candidates with target using a heatmap.

Identified features containing the most mutual information related to the target.

5. Feature Engineering

- Manually encoded candidate features before feeding into model. Utilized one-hot encoding to isolate (84) possible new training features derived by one-hot encoding (34) final feature candidates. Selected the top (10) candidates from this list of (84) one-hot encoded candidates based on features with the highest Mutual Information (MI) scores. Due to poor prediction results of model trained on these features, this approach was discarded in favor of allowing the model to encode all features during training.
- Hasty selection made of (7) new features based on schema and documentation.
- Features selected: visitNumber, totalsVisits, totalsPageviews, totalsTimeOnSite, hitsType, hitsHitNumber, and hitIsInteraction.

6. Model Building

- Models tested: BigQuery ML XGBoost Boosted Tree Classifier. Other models not attempted due to time constraint.
- Training process: Custom data split into 80% training, 10% testing, 10% evaluation.
- Evaluation metrics: Precision, Recall, Accuracy, F1 Score, Log Loss, and Receiver Operating Characteristic - Area Under the Curve (AUC-ROC).

7. Model Evaluation

Model Name: add_to_cart_xgb_a1

Model Performance on initial (10) engineered features:

- Precision: 0.861173
- Recall: 1.0
- Accuracy: 0.994219
- F1_score: 0.925409
- Log_loss: 0.298622
- ROC-AUC: 1.0

Model Name: add_to_cart_xgb_b1

Model Performance on final (7) raw features:

- Precision: 0.221602
- Recall: 0.999132
- Accuracy: 0.875443
- F1_score: 0.362748

- Log_loss: 0.277895
- ROC-AUC: 0.948758

8. Hyperparameter Tuning

Both models "add_to_cart_xgb_a1", "add_to_cart_xgb_b1" trained with the following hyperparameter tunings:

- num_trials=8,
- max_parallel_trials=4,
- HPARAM_TUNING_OBJECTIVES=["roc_auc"],
- EARLY_STOP=True,
- LEARN_RATE=HPARAM_RANGE(0.01, 0.1),
- MAX_TREE_DEPTH=HPARAM_CANDIDATES([5,6]),
- AUTO_CLASS_WEIGHTS = True

9. Results and Discussion

Interpretation of model results.

Insights: A few raw features selected in haste performed better than those identified from selective elimination and one-hot encoding.

Limitations: Dataset complexity and project time constraint.

10. Conclusion

Key findings: The categorical "hitsType" feature contributed the most to model performance. A similar, small set of features intuitively identified from the study of the documentation and existing no more than one layer deep in the dataset schema will likely improve model performance.

Recommendation: Train and test model with a few simple, and intutive features. Test other models and compare performance. Avoid totals features as they will not be available at time model is asked to make predictions.

Appendix: Performance Monitoring & Maintenance Plan

Utilize BigQuery monitoring with Google Cloud Monitoring, Google Cloud metrics, along with performance metrics queries utilizing Monitoring Query Language(MQL) in order to conduct performance monitoring and implement a performance maintenance plan.