Adludio DE/ML Challenge

Data transformation and KPI prediction

The Data

- The dataset is compiled at different stages of the adludio projects.
- At the start of the project, Adludio collects as much data as possible from clients, so that a more satisfactory Ads are designed and a successful campaign is conducted. From the analysis, I have discovered that CPEs and campaign objectives are one of the most important variables for KPI prediction.
- The second dataset is the campaign dataset which is collected during the actual Inventory bidding and Ad rendering time. Browser and site name are the most important variables discovered here. And fox news is site is the most deployed area.
- The last dataset is generated out of the creatives themselves using computer vision techniques. Although the dataset was quite limited in number compared to the above two, I discovered that there are some labels and texts that are more associated with first_dropped events than just impression. Please check the word clouds in my EDA

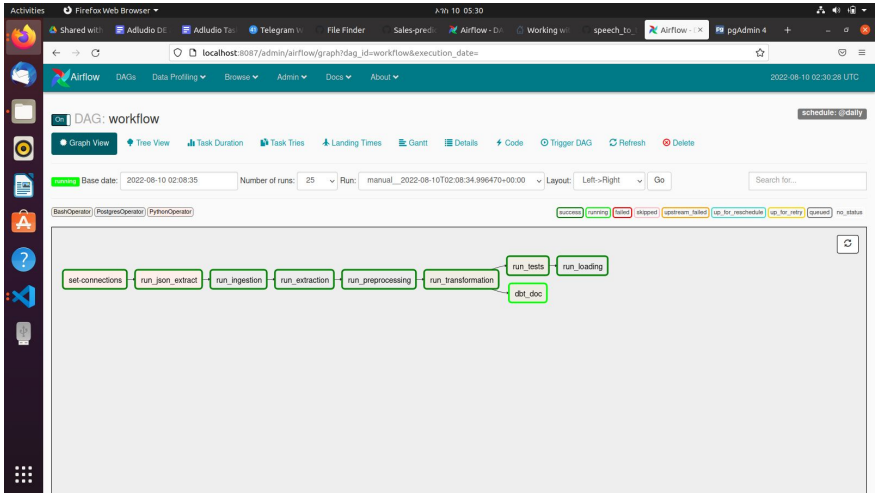
ML Algorithm Choice

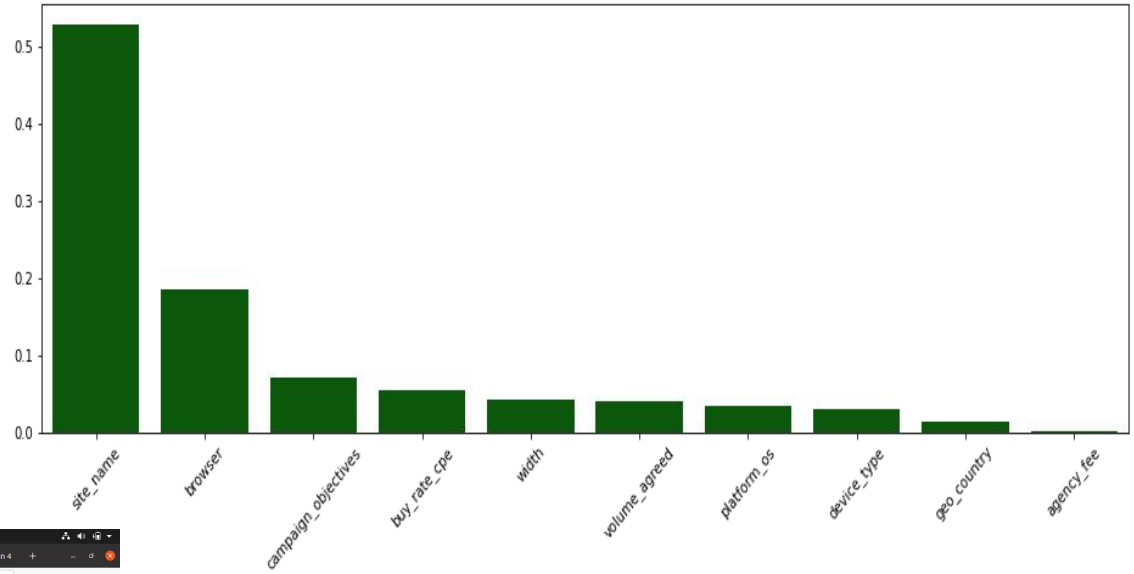
I have selected Random Forest Classifier over Linear regression for the following reasons

- Random selection in individual decision trees of RFC can capture more complex feature patterns to provide the best accuracy. Although there is the risk of overfitting, ther risk can be mitigated with the large size of the dataset.
- RFC allows me to take higher order interactions into account (not just linear). But this has some sacrifice with interpretability.
- Last but not least is that the dataset is categorical variables dominant, So If the dataset has more Categorical data and consists of outliers it is better to use Random Forest Classifier.

Results

The Data Pipeline Result





The Machine Learning Result

Thank you! 😀