Logo, company name

Description automatically generated

**Project1: Orange Data Analysis**

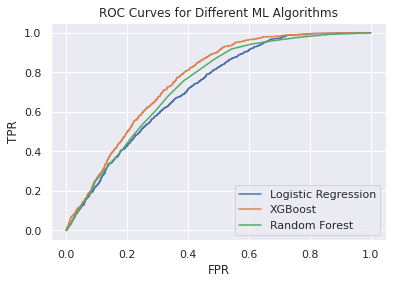
**Dikshant Agarwal, Rampunit Kumar, Yash Surange, Pratyush Pattnaik**

**Model Training and Results**

Target Variable proportion with positive flag is only 7.6%, thus it’s highly imbalanced classification problem. For this dataset, we tried various models such as Logistic Regression, XGBoost and Random Forest. Finally, based on AUC as the deciding metric, we selected XGBoost as the final model.

Table

Description automatically generated

**** Chart, treemap chart

Description automatically generated

A picture containing chart

Description automatically generatedTable

Description automatically generated

Table

Description automatically generated

**Motivation and About the Project**

**Motivation:** Customer Relationship Management (CRM) is a vital part of the new marketing and growing world. The French telecom company Orange, works in the field of telecommunication and internet service providers. One of the important aspects to drive more value and increase revenues in this generally low-margin business is leveraging your existing customer base to bundle and upsell high-cost services to potential customers. Since the overall conversion percentage of customers interested in these high-cost services is generally low, using conventional marketing and promotion methods can be quite inefficient in finding the right set of customers. Thus, here we try to devise data-driven approach to target these potential customers better than traditional targeting methods.

**About the Project:** The project is to predict the sales technique whereby a salesman attempts to have the customer purchase more expensive items, upgrades, or other add-ons in an attempt to make a more profitable sale. Up-selling usually involves marketing more profitable services or products, but up-selling can also be simply exposing the customer to other options he or she may not have considered previously.

**Conclusion:**

If we used traditional approach of random calling the customers in test set, we would’ve obtained avg conversion percentage of 7.6%. In comparison, the final model obtained by this project can help us strategize the promotions better than this traditional approach. We can divide these customers in different buckets and run differential promotions to target these customers better. Also, the probability scores obtained using final XGBoost model can help us prioritize our calling efforts.

**Future Work:**

We can further improve upon this model through following ways:

1. By improving on the pre-processing done on numerical and categorical variables, since it had a huge impact on the performance during our limited trials. We can do more context appropriate imputation for missing values.
2. Since it’s an highly imbalanced class problem, we can try variety of sampling methods to improve target variable proportion prior to modelling.

Classification Report for final XGBoost Model

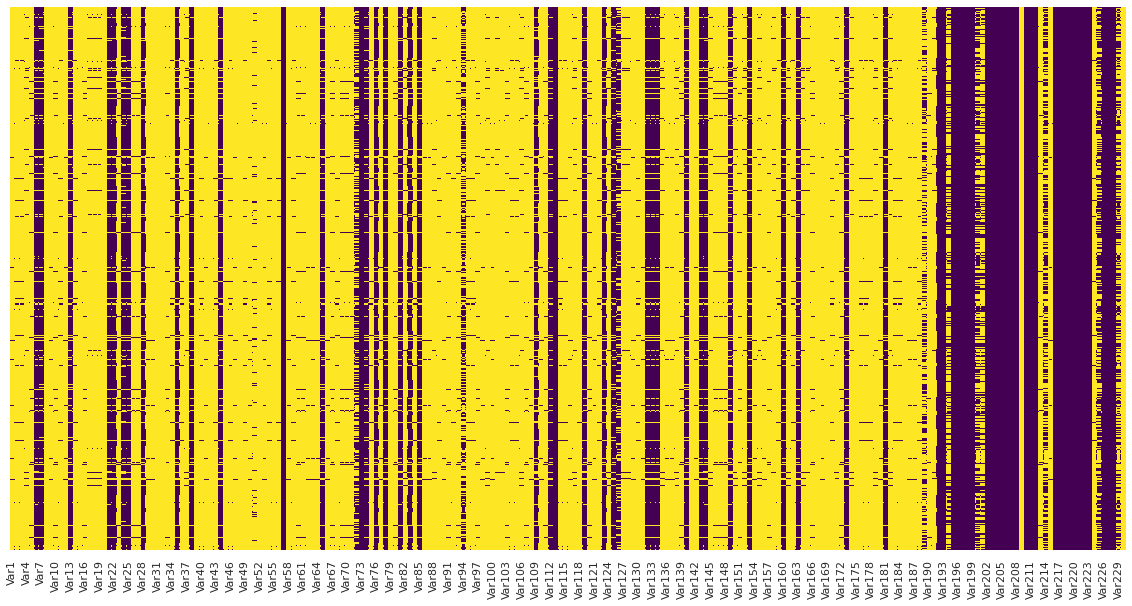
Bucket-wise Conversion Percentage for final XGBoost Model

Feature Importance Plot for final XGBoost Model

Confusion Matrix for final XGBoost Model

**Data and Labels**

Raw data consists of 50000 rows and 230 columns with Nan values, after cleaning, filtered to 50000\*66, consists of categorical variables and numeric variable as response and categorical variable as predictor.

****

**Figure: Heat map showing nan values in the dataset, where blue color showing the actual values and yellow color showing the Nan values.**

**References**

1. https://medium.com/@pathakvishnudutt123/customer-relationship-prediction-kdd-cup-2009-2248d83e9d32

2https://github.com/yulguseva/orange\_churn/blob/master/Submission.ipynb

3 https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/

4. http://www.mtome.com/Publications/CiML/CiML-v3-book.pdf?ref=https://githubhelp.com

5. https://github.com/Wangsherpa/Customer-Relationship-Prediction/blob/main/KDD\_2009\_data\_preprocessing.ipynb?ref=https://githubhelp.com.