Data Modelling

- ONLINE PURCHASE INTENTION -

By: Tanvi Gandhi

• PRE-PROCESSED DATA

Before applying various Machine Learning techniques, we encoded and scaled the data. The 7 categorical features were encoded using *get_dummies* function with the first column dropped for each feature to avoid correlation. This resulted in total of 68 columns including numerical and categorical features. The target variable *Revenue* was label encoded.

The data was further feature scaled using *StandardScalar* function. It was then split into 80% training set and rest testing set.

• NAÏVE BAYES

The typical Gaussian Naïve Bayes for the given dataset gave just 29% accuracy which is very low for our scenario. We then balanced the data using SMOTE algorithm and the accuracy increased to 63%.

• LOGISTIC REGRESSION

For Logistic Regression also, we performed both on the original dataset and balanced dataset using SMOTE. The accuracy for former method is 88% and with latter strategy is 86%. It is to note that the balanced data has a reduced accuracy.

• RANDOM FOREST

For Random Forest, we tried 4 variants. The first is, Random Forest with given dataset which gave 89% accuracy. Next is Random Forest with 40 important extracted features out of 68 and gave the accuracy of 90.51%. The third is, Random Forest with given dataset and balancing using SMOTE which gave 94% accuracy. The last variant we tried is Random Forest with 50 extracted important features and balancing using SMOTE. It gave the accuracy of 90.4%.

- XGBOOST

We ran the default XGBoost algorithm which gave us 90.51% accuracy. We then tried to fine tune the parameters and the accuracy was 90.1% suggesting that we need to further hyper tune the model.

- AGGREGATED ANALYSIS

The below table shows that Random Forest with extracted features and default XGBoost gave approximately equal accuracies and Random Forest with SMOTE gave highest accuracy of all other models.

Model Variation	Accuracy
1. Naïve Bayes	29%
2. Naïve Bayes with SMOTE	63%
3. Logistic Regression	88%
4. Logistic Regression with SMOTE	86%
5. Random Forest	89%
6. Random Forest with important features	90.51%
7. Random Forest with SMOTE	**94%**
8. Random Forest with important features and SMOTE	90.4%
9. Default XGBoost	90.51%
10. Tuned XGBoost	90.1%