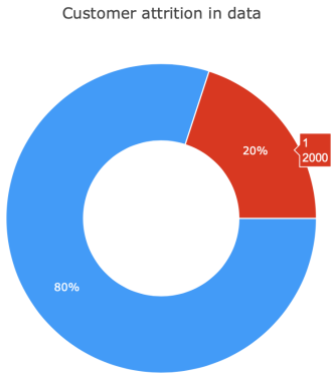
Advertising Churn Prediction

Non-technical report

Problem statement - As a media company, helping local small and medium sized business through digital advertising is one of our main revenue streams. We do so by running advertising campaigns on behalf of our clients on platforms such as Google and Facebook. In this exercise the task is to predict if a client will stop running advertising campaigns(churn).



The data provided to me contained 10,000 rows and 10 columns. Customer churn in the dataset was at 20% (2000 data points). The main steps that I followed for accomplishing this challenge was as follows –

1. Research about customer churn modeling
2. Data preprocessing
3. Exploratory Data Analysis
4. Data preparation for modeling
5. Model creation
6. Model comparison and coefficient analysis

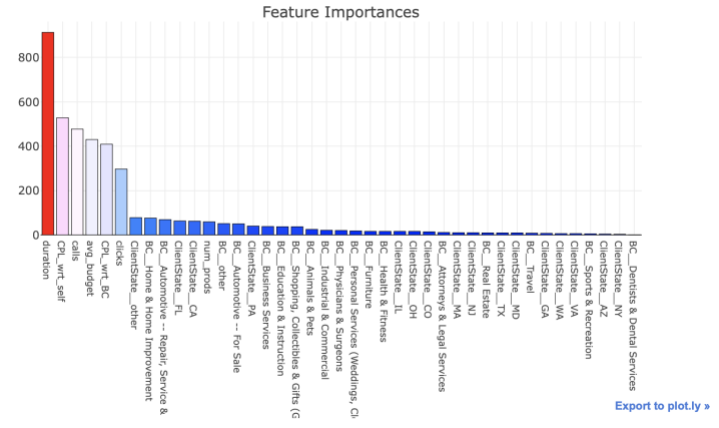
Few major overview points from the data were –

1. California, Texas, Florida and NY were the top states were most of the customers were from.
2. 26.1% of the customers were under the business category Home & Home improvement. In comparison the 2nd best shopping, collectibles and gifts just had 7.36% share.
3. When I looked into the client state distribution among churned and non-churned customers, I found out that the percentage of churned customers is more in smaller states (marked as others) and Florida. In comparison states like CA and TX has less percentage of churned customers in comparison. I would like to point out that the concentration on larger markets is indirectly hurting the churning in 'other states'
4. A similar but minor effect can be observed from business category as well. But here, Shopping, collectibles and gifts customers seems to be more happier in comparison to Automotive – Sale businesses.

In-order to model this advertising dataset, I went on to implement 6 different Machine learning algorithms and the results overwhelmingly points to XGBoost Classifier and LGBM Classifier as the best performing models.

Both the models showed same first 6 features as the most important in determining customer churn. These were –

1. duration of time the business has been with the company (almost twice as important as the next one in this list)
2. Cost per lead change in last three months
3. Number of calls
4. Average budget
5. Cost per lead change in category
6. number of clicks.



Future analysis would revolve around understanding to what degree does the important features impact customer churn and in which direction. Also, using interactions among features will allow the company to understand how customer churn get affected when more than one variable is in play.

Advertising Churn Prediction

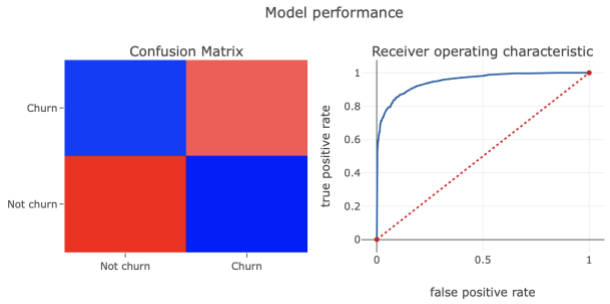
Technical report

The customer churn dataset contained 10,000 rows and 10 columns including the dependent feature. In-order to build a predictive model I decided to follow these major steps –

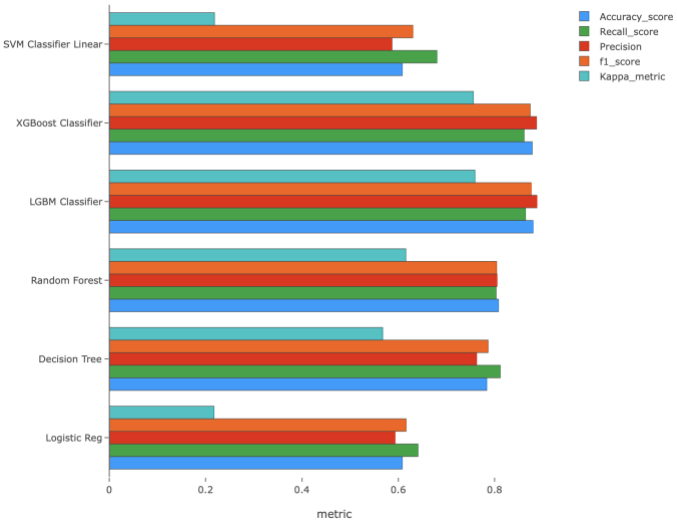
1. Research about customer churn modeling
2. Data preprocessing
3. Exploratory Data Analysis
4. Data preparation for modeling
5. Model creation
6. Model comparison and coefficient analysis

First, I read about customer churning and other usual machine learning methods used to predict similar problem statements. During the initial data preprocessing steps, I got a high-level overview of the dataset. For example, I got to know that the customer churn in the dataset was at 20% (2000 data points). Exploratory Data Analysis was done to generate graphs for numerical and categorical feature distribution among churned and non-churned customer population. Correlation plot was also generated to check for correlated features. I also checked the frequency distribution for categorical features and possible outlier existence for numerical features. For client_state and BC I decided to reduce the number of categories by merging low frequency points into 'other' category. 1092 missing values were spotted in the CPL_wrt_self feature space. I decided to remove them instead of filling them with 0's, mean, or median as I believed the remaining 9000 data points were enough for building a good model. Also, if originally a pattern existed between the missing values, then they could end up skewing the final model. To tackle the significant imbalance in the dataset with 80% non-churned customers and 20% churned customers I carried out SMOTE (Synthetic Minority Over-Sampling Technique) sampling method. I also normalized the numerical features before passing them into each of the models.

For modeling, I picked 6 different Machine learning models. These were –
1. Logistic Regression
2. RandomForest Classifier
3. Decision Tree Classifier
4. LGBM Classifier – The image on the right shows the confusion matrix and the ROC curve for LGBM.
5. XGBoost Classifier
6. Support Vector Classifier



LGBM and XGBoost Classifier performed the best. I used these algorithms as I knew they generally perform well on structured datasets. In-order to determine model performances I used metrics such as Accuracy, Recall, Precision, F1 score and Kappa metric. Logistic Regression ended up performing the worst. Special attention was given to recall and precision to make sure there isn't any type 1 or type 2 errors. I also reduced the learning rates and depth on Tree classifiers to make sure that the trees don't overfit. Hyperparameter tuning was done manually. I also tried non-linear spline model using SVM but it didn't improve the overall performance.



Future work to improve the current model would be to carry out step wise selection, add interactions between variables. Also, using interactions among features will allow us to get a deeper understanding on variable relationships and how they contribute to the accurate prediction of churning.