Machine Learning Project - Churn Modelling

Business Objective

Customer churn is a concerning problem for large companies (especially in the Telecom field) due to its direct effect on the revenues. Companies often seek to know which customers are likely to churn in the recent future so that a timely action could be taken to prevent it.

Problem Statement

You have been assigned the task to build for this Telecom company a Logistic Regression Machine Learning model that predicts which of their customers are likely to churn (stop using their service in future). For more details on the intermediate and final outputs expected, refer to the list of deliverables mentioned in the “Model Building” and “Model Validation” sections below.

Tools: Python

Data Description

The dataset provided for this activity consists of 11 features where 10 are independent features and 1 is a target variable. Features in this dataset are described as below:

- Churn: 1 if customer cancelled service, 0 if not (Target)
- AccountWeeks: number of weeks customer has had active account
- ContractRenewal: 1 if customer recently renewed contract, 0 if not
- DataPlan: 1 if customer has data plan, 0 if not
- DataUsage: gigabytes of monthly data usage
- CustServCalls: number of calls into customer service
- DayMins: average daytime minutes per month
- DayCalls: average number of daytime calls
- MonthlyCharge: average monthly bill
- OverageFee: largest overage fee in last 12 months
- RoamMins: average number of roaming minutes

Model Building:

Task - 1

- Show Bi-variate plots (scatter/ bar) of all meaningful variables with the dependent variable.
- Present your final model results (show what’s applicable from the below list):
- List of variables which came significant.
- Beta coefficients for the respective variables along with, p-values

- In case of a Decision Tree model, show Rules or the Decision Tree
- Summarize the steps followed to finalize your model - consisting of the below steps (as applicable)
- Sampling
- Feature Engineering
- Performance comparison between Train and Test
- Use of Cross validation
- While developing the model, you would have gotten a few candidate models which were not as good as the final model (in terms of performance, multicollinearity, or statistical stability etc.). Show a few of these candidate models and explain their shortcomings.
- Show what kind of feature engineering did you apply in your project and why (include in your results what's applicable from below)
- Dummy Variables
- Label Encoding
- Any bin-based variable created -what was the significance/rationale of binning
- Any new derived variables created using the raw variables – For e.g., Ratio based, difference based, % difference based / Rate of change, etc.
- If the provided dataset is unbalanced, what steps did you take to balance? Also explain the technique used to oversample/undersample the dataset? (Assume an event rate of less than 5% indicative of class imbalance, you can still use oversample/undersample to improve model performance even if the event rate is between 5-10%)

Task - 2

- Demonstrate Live how your model will assign class/ or compute the probability for a new data point?
- Provide your understanding of the next steps that the client/ end user needs to follow to deploy your model at their end. Think on the below lines:
- Any technical/infrastructure requirements that the client needs to meet?
- What files do you need to provide them?
- What kind of data cleaning and preprocessing would the client need to do before using the model?
- How will the client use your model on new data?
- How will the client know that the model is performing well on new data points?
- Show your model's performance on the below metrics (both train and test samples)
- Confusion Matrix
- Classification Report
- Rank ordering test results (Rank ordering is an important measure of model performance and its ability to separate out the event from the non-events). You can divide your probability predictions into deciles and look at the event rate in each decile.
- For the given business problem which of the below metric(s) did you choose and why? Include in your final output any additional activity performed (and its results) to

get to the best values of the below metrics (F1-Score, AUC-ROC curve, AUC-ROC Accuracy). Accuracy

- Accuracy
- Precision
- Recall

Deliverables

- A well-designed deck outlining the conclusions and the analysis (.ppt)
- A well-structured code pushed on GitHub (Write an informative README, well-structured code/notebooks)
- *Optional*: A blog post on medium/personal blog/blogger/LinkedIn