Customer Churn Risk Prediction

Problem Statement: - As business, we would like to understand what is the risk of customer churning. We would like to create predictive model that has output per customer, where we have KPI that indicates the churn risk.

Proposed Solutions

(a) How would you approach this modelling?

i. Considering the technical data like (Wi-Fi Quality, network performance, band-with usage) and we don't have any target predicted column like customer churn or not (i.e., 1 for churn, 0-not churn) in our data set.

ii. We can model such predictive analytics using un-supervised machine learning method.

iii. In unsupervised learning method, we can find hidden pattern or structure in the given technical data.

iv. Precisely clustering the customer data into say 4-5 cluster (say **very low churn risk**, **low churn risk**, **medium churn risk**, **high churn risk** and **very high churn risk**) based on above technical data.

(b) How can possible model look like?

i. Since model needs to processes 6M customer data in TB and that too for 30 days period so this requires distributed processing mechanism on Hadoop cluster.

ii. We can partition the data on spark worker nodes and K-Mean (i.e. k=5) clustering algorithm can be run on each worker node parallelly. Master node can share the initial/default cluster centroids (say c1, c2, c3, c4 & c5) info at the beginning to each worker node. In first iteration clustering algorithm will assign data points (w/ above technical features w.r.t customer) to respective centroid.

iii. After each iteration, Master node can collect and store count of data points associated to each centroid (c1, c2, ...c5) from all worker nodes and store it in memory for future reference.

iv. Master node now can compute updated centroids for each c1 to c5 centroid set received from worker node in step(iii) and new centroid set c1 to c5 will be shared with respective worker node.

v. This process keeps iterating till convergence criteria is met. Once a convergence criterion is met Master node process collects local clusters and combines them into a global one.

vi. At last, we can have 5 cluster and based on their feature values for which they group together, we can assign above 5 tag on graph (very low churn risk, low churn risk, medium churn risk, high churn risk and very high churn risk)

.

Customer Churn Risk Prediction

(c) What should we consider such problem?

i. Like in typical clustering problem, we should consider scaling all the data points on Standard Scale for all variables (i.e., technical feature) so convergence criteria met quickly.

ii. Spark ML pipeline can be used to define pre-processing STAGES before applying algorithm.

(d) How will you deploy this model? What can be a possible deployment of this?

I. Tools like “mlflow” or amazon “sagemaker” can be used for spark model deployment. “mlflow” can be used for both local and production model deployment with UI interface to keep track of hyperparameter and model evaluations metrics.

II. MLeap tool can also be used with Amazon EMR