Domain Oriented
Telecom Churn Case Study
2023

Tariq Khan

Anshuman Barthakur

Rubina D’souza

Problem Statement

- In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate.
- Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, customer retention has now become even more important than customer acquisition.
- For many incumbent operators, retaining high profitable customers is the number one business goal. To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.
- Thus, churn prediction is usually more critical (and non-trivial) for prepaid customers than post paid customers. Also, prepaid is the most common model in India and Southeast Asia for which we are going to do the analysis.

Business Goal

- The business objective is to **predict the churn** of the **high value customers** in the last (i.e. the ninth) month using the data (features) from the first three months.

- **High Value Customers** are the customers who are in and above 70th percentile of the average recharge amount in the first two months (the good phase).

Problem Solving Methodology

- Source Data for analysis
- Data Pre-processing: Data cleaning, Data Manipulation
- EDA
- Feature Selection
- Model Building: Logistic regression Model
- Model Training
- Model Evaluation
- Model Performance
- Predictions
- Conclusion

Model Building

Base Model - **LOGISTIC REGRESSION WITHOUT ANY TUNING**

- For the problem statement, precision and recall are the important metrics.
- The base model is evaluated on train using different metrics.
- It has a very low precision and recall.
- This is very likely due to data imbalance

MODEL EVALUATION - TRAIN SET USING CROSS VALIDATION

- We can use any of the "scoring" while doing cross validation in the context of the problem at hand, we are using "precision", "recall" and "roc_auc" as the scoring

- Cross validation gives much more realistic performance evaluation of the model on train data, therefore the cross-validation score uses accuracy as the default metric to evaluate the model
- The evaluation matrices with cross validation is smaller than that without cross validation.
- Going forward we shall do model evaluation only with cross validation.

MODEL EVALUATION ON TRAIN SET

- Observations - The model has over fitted the train data

- Cross validation score gives a far more reliable estimate of the generalized performance on unseen data

- OOB Score in RandomForest is somewhat similar to cross val score

LOGISTICE REGRESSION MODEL - MANUAL RFE

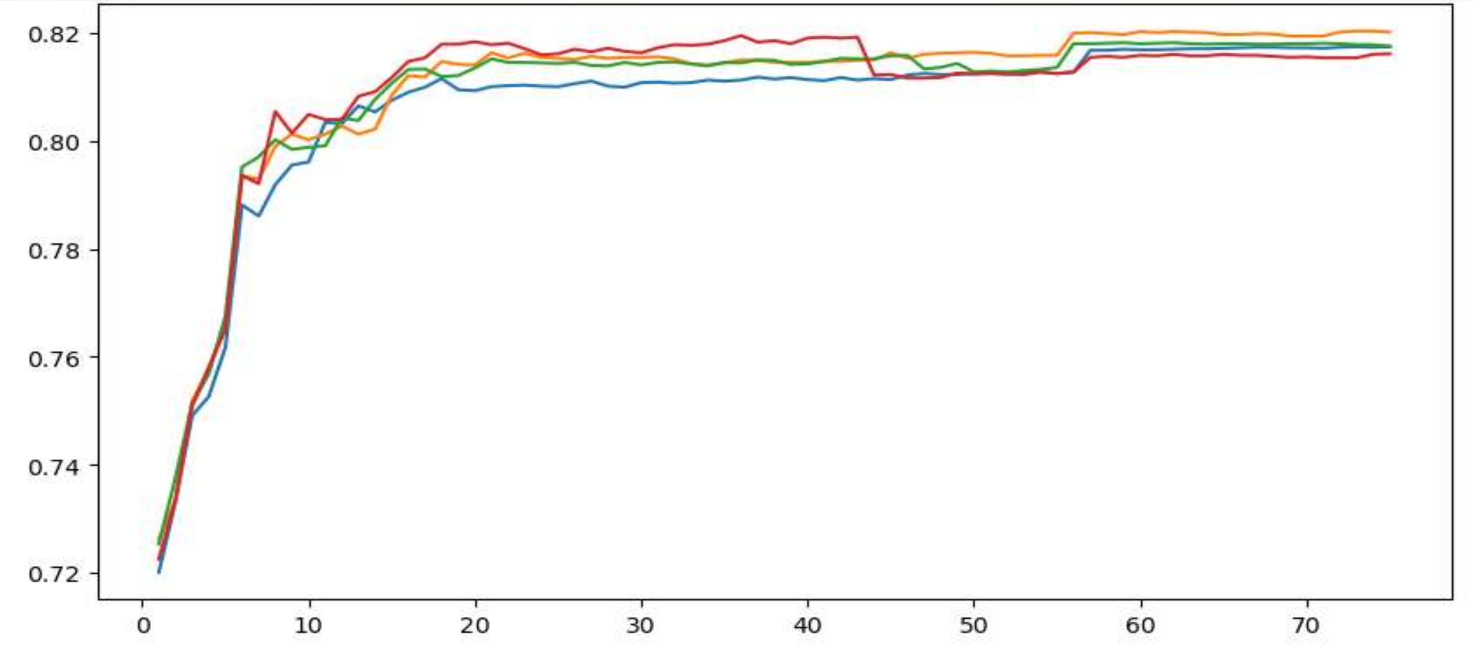
Recursive Feature Elimination - RFE

- We used random number of 10 estimators. Looks like 10 is not the optimum estimator and as a result the validation result dropped.
- We shall use cross validation feature selection - **rfecv**, which is also much faster.

Cross validation for feature selection

- This above is a manual process and for loop takes a long time if the number of features is high.
- We can automate this and get the same result by using **rfecv** which is also much faster
-

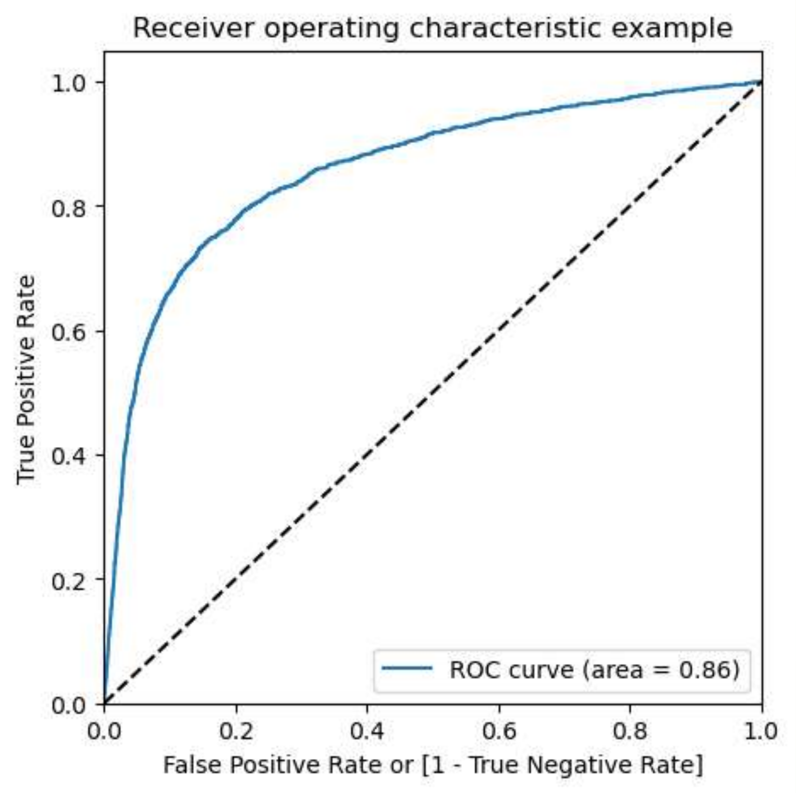
Using RFECV

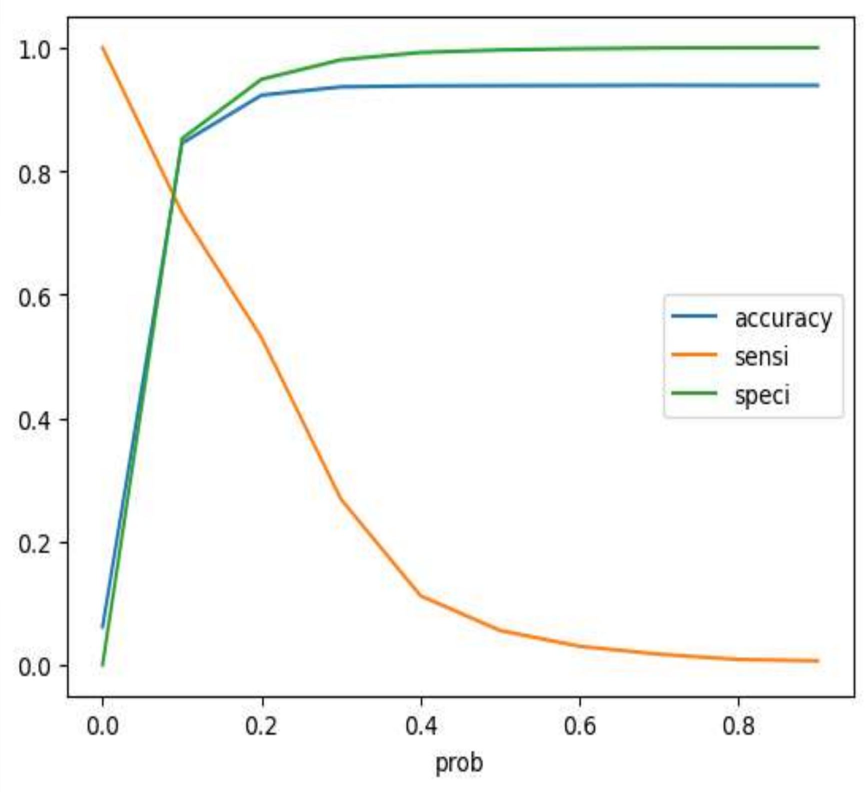


# Findings 1 - With about 15 estimators, we can get an optimized logistic regression model

# Findings 2 - With about 41-45 estimators, we can get the best logistic regression model

PLOTTING ROC CURVE





0.1 is the optimum point to take it as a cutoff probability.

#Observation - We have now got a model with high recall value which is the requirement.

PRECISION RECALL TRADEOFF

- We shall not go with 0.3 threshold, since this will reduce the recall value.
- As retaining existing customers is cheaper, we are okay with giving away some offers to existing customers who are actually not going to churn

THE MOST IMPORTANT PARAMETERS ARE -

- total_ic_mou_8
- total_ic_mou_7
- total_rech_amt_8
- avg_rech_amt_6_7
- vol_3g_mb_8
- monthly_3g_8
- loc_og_mou_8
- last_day_rch_amt_8
- monthly_2g_8
- roam_og_mou_8
- sachet_2g_8
- spl_ic_mou_8
- THRESHOLD VALUE IS 0.1

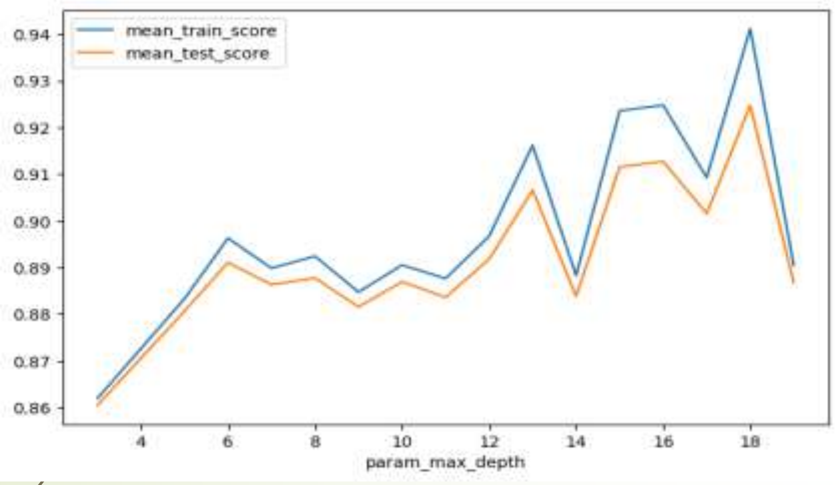
RANDOM FOREST
Hyper-parameter tuning using Cross Validation

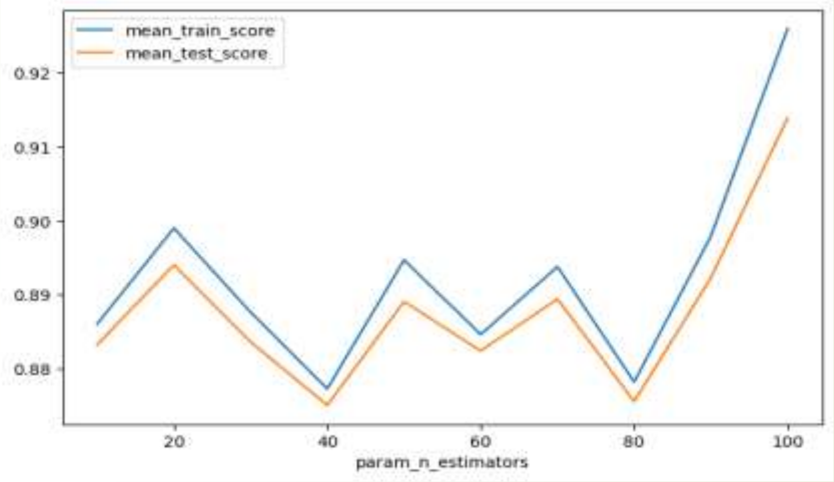
- Randomized searchcv is a highly efficient technique that is used to identify the best set of hyper parameter values in a fewer number of iterations.
- This technique performs quite well at a reduced cost and a shorter time for huge data sets and models with large numbers of hyper parameters.
- Randomized search cv is similar to grid search cv but randomly takes samples of parameter combinations from all possible parameter combinations.

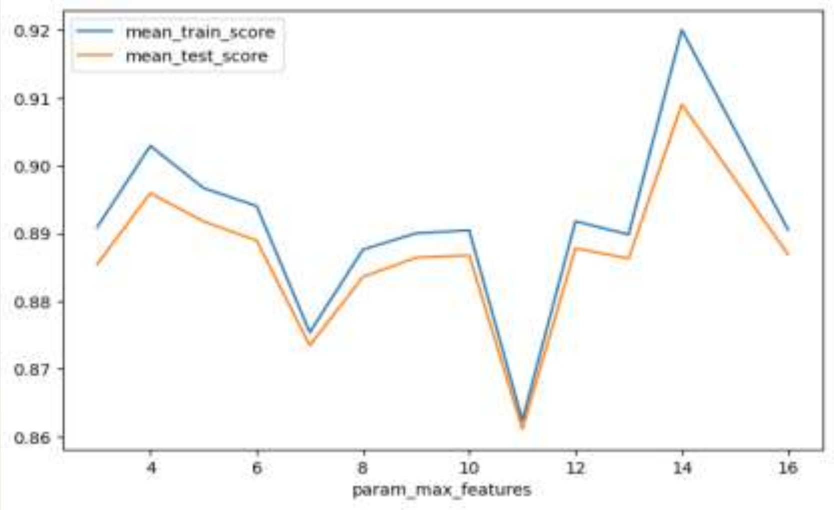
INFERENCE:

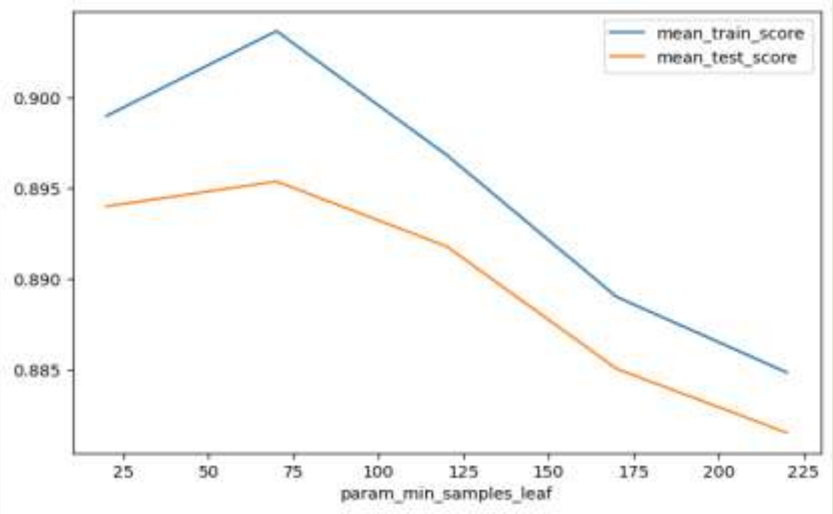
- The difference in performance of the top 20 models are very similar. 0.94xxx. If we want a decent model considering also the computational power, we can consider any model in the list. For the current case, we shall move ahead with the best model.

Effect of Hyper-parameter









MODEL EVALUATION

EVALUATION OF LOGISTIC REGRESSION MODEL:

Observation - The logistic regression model worked equally well on train and test data.

EVALUATION OF RANDOM FOREST MODEL

Logistic regression gives a better recall value, therefore, finally considered model is Logistic regression.

Conclusion

- Telecom Company needs to pay attention to the roaming rates. They need to provide good offers to the customers who are using services from a roaming zone.
- The company needs to focus on the STD and ISD rates. Perhaps, the rates are too high. Provide them with some kind of STD and ISD packages.
- To look into both of the issues stated above, it is desired that the telecom company collects customer query and complaint data and work on their services according to the needs of customers.
- Given our business problem, to retain their customers, we need higher recall.
- As giving an offer to an user not going to churn will cost less as compared to losing a customer and bring new customer, we need to have high rate of correctly identifying the true positives, hence recall.
- Provide offers to customers who has a churn probability of more than 10%
- For times when churning is high, company can target even smaller threshold e.g. 8%.