COMPARISON AND CONTRAST OF POINT & CLICK, AUTOML AND PROGRAMMING TOOLS IN PREDICTING CUSTOMER CHURN IN TELECOMMUNICATION INDUSTRY

1.0 INTRODUCTION

Customer churn is usually one of the major problems in many businesses. Previous studies have shown the fact that attracting new customers is much more expensive than retaining the existing ones. Therefore, companies focus on building an accurate and reliable predictive model to identify customers that would churn soon. Looking at the telecommunication industry where for example we can predict customer who are not going to renew their purchase annually, monthly, weekly or daily using the various ML tools. This project would access analytic tools such as Point & Click (SAS Viya and Rapidminer), AutoML (DataRobot) and Programming (Python) in predicting customer churn.

1.1 DATASET (SAS Viya)

The commsdata, which contains information about customer behaviour. The input variables include demographic information, product usage and type, billing data, and interactions with customer service. There is also a binary variable that indicates whether customers churned, and this is the target to predict. The commsdata data set contains 128 columns and more than 56,000 rows. Raw Data access**:** commsdata.sas7bdat

1.2 DATASET (RapidMiner, DataRobot and Python)

Telcom customer churn data from Kaggle with raw data containing 7043 rows (customers) and 21 columns (features). The churn was set as our dependent or target variable. Raw Data access**:** Telecom Churn Prediction | Kaggle

2.0 POINT AND CLICK (SAS Viya)

SAS Viya is a cloud-enabled point and click tool, in-memory analytics engine that delivers everything with a quick, accurate, and consistent results. SAS Viya is mostly used by businesses to explore and address complex analytical challenges of today and effortlessly scales to meet future needs. SAS Viya passes through three phases of the analytic life cycle which are data, discovery, and deployment.

The model studio for SAS Viya was used for this analysis. This dataset was partitioned into training and validation data where the training set was used to build the model and the validation dataset to assess the performance of the data to find the sweet spot between bias and variance. The prediction of **churn** (binary target) is the main goal using this various inputs variables and train supervised learning models for our prediction. The pipeline as seen in pipeline fig 1.0, the data passed through various nodes to get our model comparison and prediction. First the replacement node was used to replace outliers with specific value. Second, the transformation node was used to apply numerical transformation to the input variables. Third, added new features with the text mining node using the text variable **verbatims,** which is one of five text variables in the **commsdata** data source, hence rejecting the other four variables**.** Fourth, the variable selection nodes to reduce the number of inputs for modelling before we started making model comparison.

2.1 Project Summary and insights

The champion model for this project is Gradient Boosting as seen in fig 1.3. The model was chosen based on the KS (Youden) for the validate partition (0.59) 93.86% of the validate partition was correctly classified using the Gradient Boosting model. The five most important factors are Handset Age Group, Total Days Over Plan, Days Suspended Last 6M, Transformed MB of Data Usage Month 6, and Total Late Payments Lifetime as seen in Fig 1.2. The ROC curve of the champion model is also displayed in fig 1.1**.**

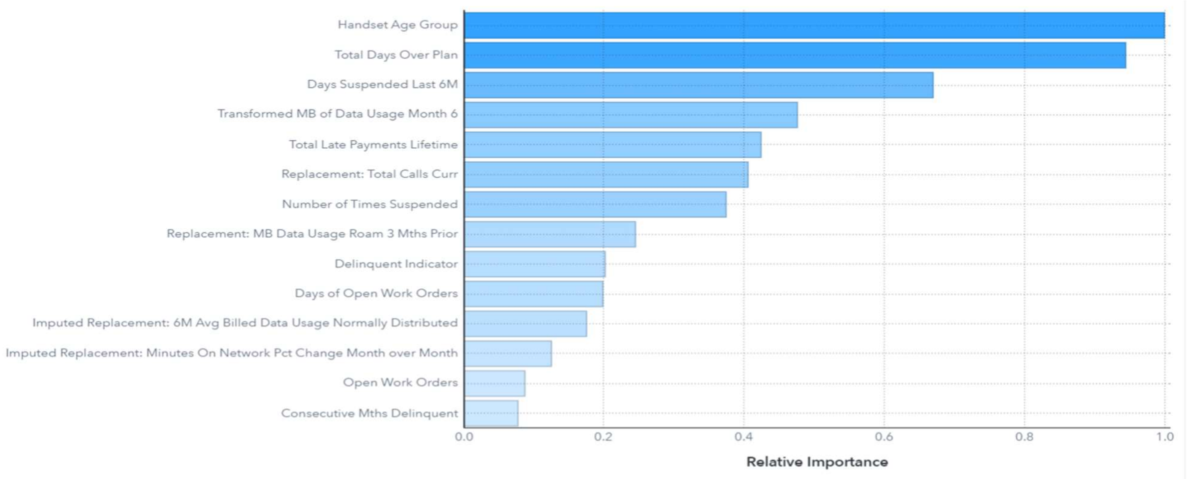


Fig 1.2 Most Important Variables for Gradient Boosting model (Champion Model)

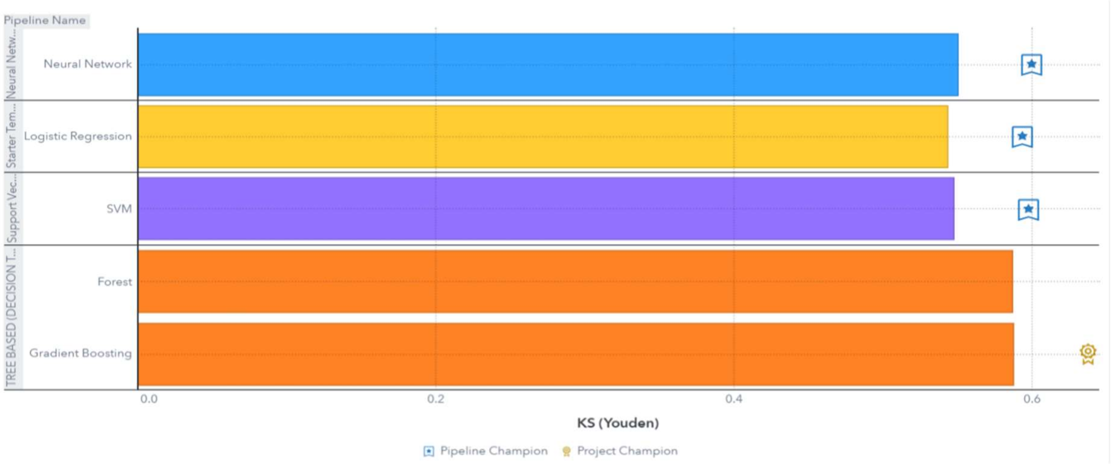


Fig 1.3 Assessment for All Models

RapidMiner

RapidMiner which is also a point and click tool is very similar to SAS Viya. The RapidMiner studio gives us a design view as work areas to carry out the telcom customer churn prediction task. Looking at fig 1.4 we could see the pipeline the Telecom Churn data | Kaggle passed through. This parameter was useful to create decision tree as seen in fig 1.5.

3.0 AutoML (DataRobot)

DataRobot is an AutoML tool that helps data analyst and scientist to do more with less by automating the time consuming and repetitive tasks. DataRobot Pipelines enable data scientist and analyst to build and run machine learning data flows. This pipeline just like the point and click pipelines starts by collecting data from various sources, cleaning, combining them, standardize the value among other data preparation operation to build a model without writing any code unlike the use of a programming language or point and click tools. DataRotbot also provides a leader board with scoring which provides the summary of each model built in the project and a model comparison as in fig 2.2 like the model comparison in SAS.

3.1 Summary and Insight

Looking at Fig 2.1 show the Tree-based variable importance this chart shows the relative importance of each features calculated relative to the most important feature for predicting customer who would churn, hence the tenure is the most important independent variable or relative importance of all key features making up the model having a 100%, we can see some customers have been with the telecom company for just a month, while quite a many are there for about 72 months having different contract. The Light Gradient Boosted Tree classifier with early stopping on the leader board was used for our analysis where the data passed through the tree-based algorithm pipeline before making prediction as seen in fig 2.0 and also an ROC curve as seen in fig 2.1

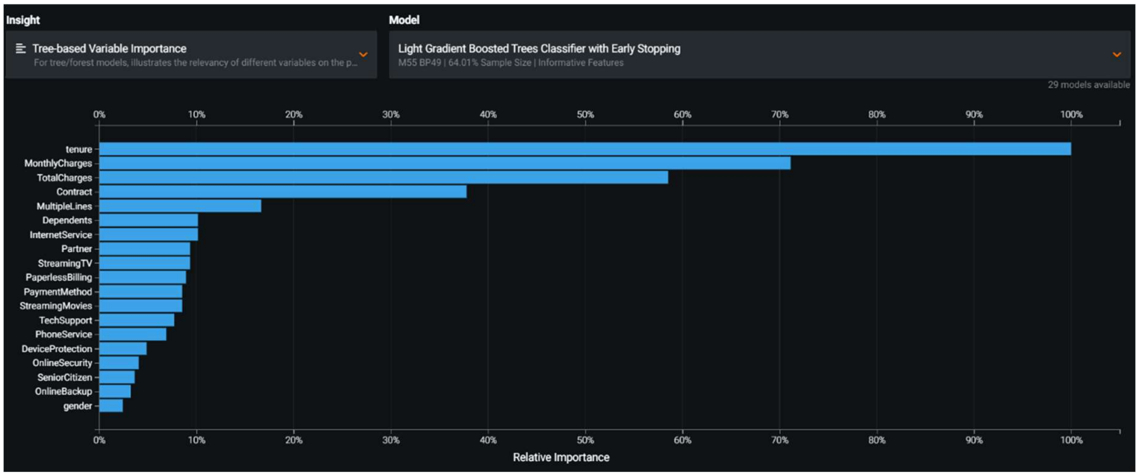
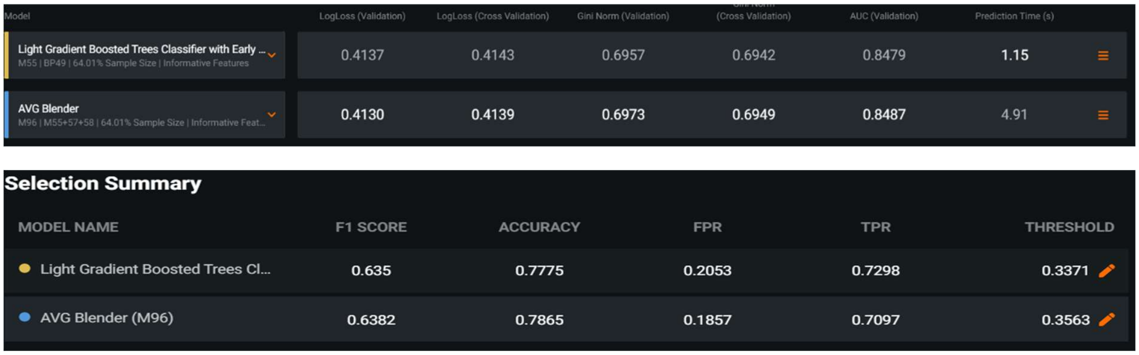


Fig 2.1 Relevance of other variables to churn

Like the point and Click tool we also have a model comparison tab to make to compare different models to know the different validation score and prediction time. As seen in Fig 2.2



Model	LogLoss (Validation)	LogLoss (Cross Validation)	Gini Norm (Validation)	Gini Norm (Cross Validation)	AUC (Validation)	Prediction Time (s)
Light Gradient Boosted Trees Classifier with Early ... M55 | BP49 | 64.01% Sample Size | Informative Features	0.4137	0.4143	0.6957	0.6942	0.8479	1.15
AVG Blender M96 | M55+57+58 | 64.01% Sample Size | Informative Feat...	0.4130	0.4139	0.6973	0.6949	0.8487	4.91

Selection Summary

MODEL NAME	F1 SCORE	ACCURACY	FPR	TPR	THRESHOLD
● Light Gradient Boosted Trees Cl...	0.635	0.7775	0.2053	0.7298	0.3371
● AVG Blender (M96)	0.6382	0.7865	0.1857	0.7097	0.3563

Fig 2.2 Model Comparison and summary

My previous work [Airbnb/DATAROBOT]that was done as a team. DataRobot was a useful tool in predicting a large time series data without writing codes or using point and click parameters.

4.0 Programming (Python)

Programming language can also be a useful tool for data analysis going through EDA, Future Engineering, Data Cleaning, Model Training, piper parameter tuning and as well model deployment by writing codes. An example using the Telco Churn Dataset, where I would reference a good prediction done by a Kaggle user "**atindrabandi"** with a published work on prediction of customer churn Telecom Churn Prediction | Kaggle using python.

Furthermore, the machine learning process using python programming tool followed this step as seen on the Kaggle notebook.

- First, libraries for machine learning were imported for analysis.
- Second, exploration of the imported data to see if there are any missing values which was detected and removed.
- Third, the predictor variable was converted to a binary numeric variable and all categorical variables to dummy variables.
- Fourth, EDA (Exploratory data analysis) was performed on demographics, Customer account information, Distribution of services used by customers, relationship between monthly and total charges and interaction of the churn variable with the independent variables in a correlation plot. This exploratory data analysis we also discovered that 74% of the customers do not churn.
- Lastly, a model comparison was developed using Logistic Regression, Random Forest, Support vector machine (SVM), ADA Boost and XG Boost. Interestingly XG Boost was the champion model with an accuracy of the test data of almost 83%.

This is the same similar steps we used in SAS model studio however, python requires the use of code and SAS VIYA do not.

5.0 Conclusion

Point & Click (SAS Viya and Rapidminer), AutoML (DataRobot) and Programming (Python) which was used in predicting the best model for customer churn had similar pipeline and process. However, the data was handled differently to produce results.

Furthermore, Point & Click (SAS Viya and Rapidminer) tools requires a data scientist to handle and adjust its parameters to get the best performing model unlike the use of AutoML which is mostly automated whereby a non-data scientist or analyst can handle and tell the best performing model for prediction or classification. Nevertheless, the use of a programming tool for machine learning requires the use of various libraries and writing codes to know the best performing model.

Conclusively, the analysis of customer churn in telecommunication while using these various tools was useful in our model comparison and prediction. AutoML (DataRobot) was fast having a prediction time of 1.15 seconds as seen in fig 2.1, however it was too automated would recommend its users to have more control of its functionalities happening behind the scenes. Point and Click (SAS and RapidMiner) was useful but with a shortcoming when it rerun its pipelines to get the best performing model. Programming (Python) needs a programmer or data analyst or scientist who understand its functionalities to code and interpret the output.

APPENDIX

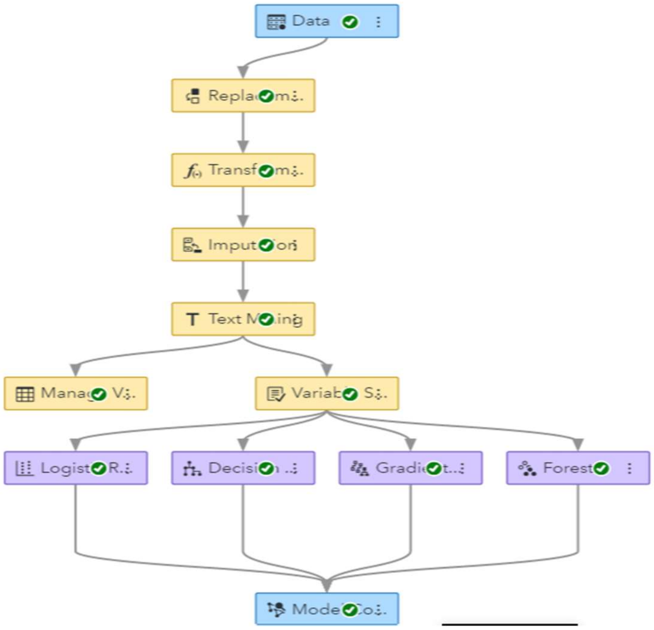


Fig 1.0 SAS Viya pipeline

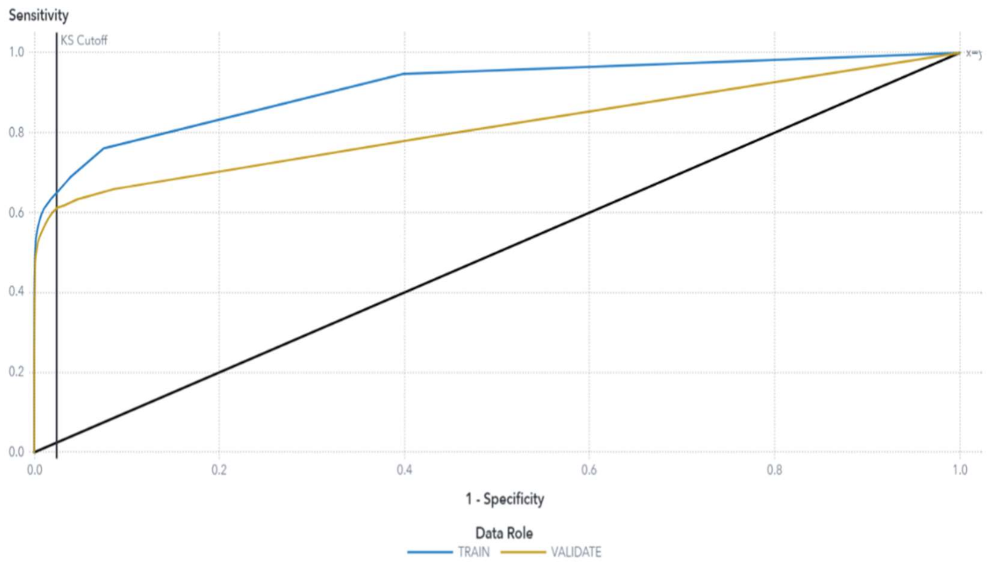


Fig 1.1 ROC Curve (SAS Viya)

Raw Data access: commsdata.sas7bdat

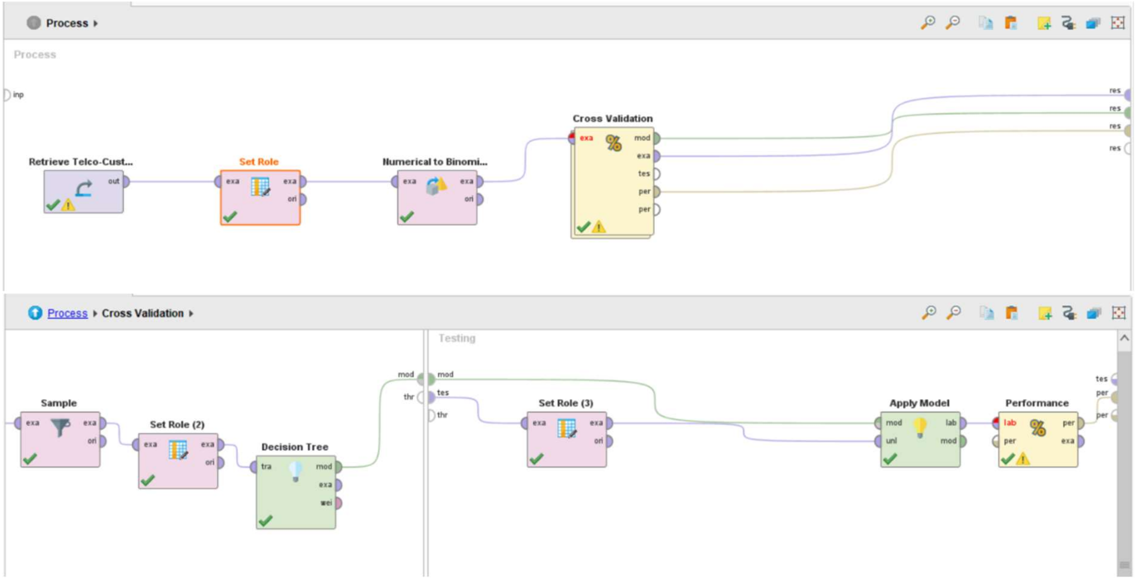


Fig 1.4 RapidMiner Pipeline

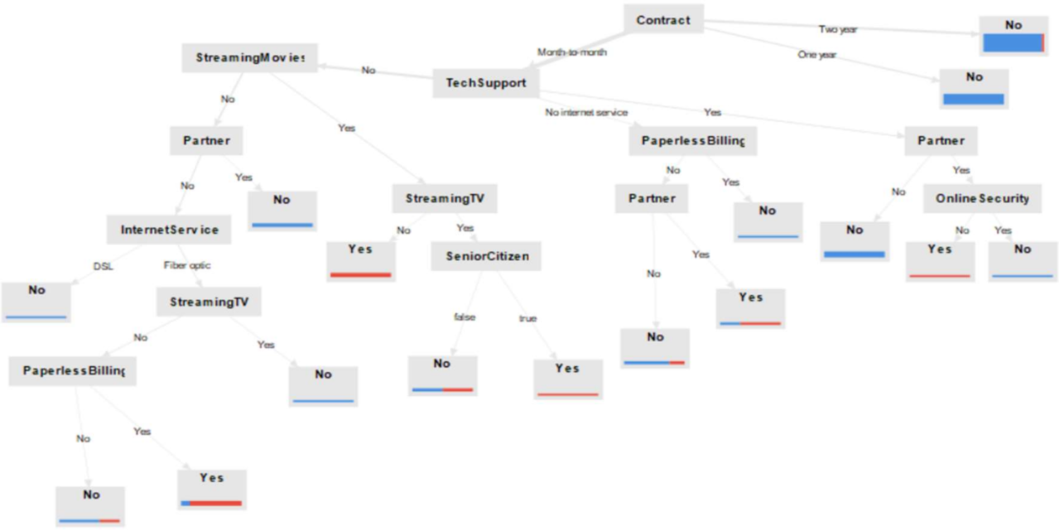


Fig 1.5 Tree (Decision Tree)

Raw Data access: **Telecom Churn Prediction | Kaggle** (https://www.kaggle.com/code/bandiatindra/telecom-churn-prediction/data?select=WA_Fn-UseC_-Telco-Customer-Churn.csv)

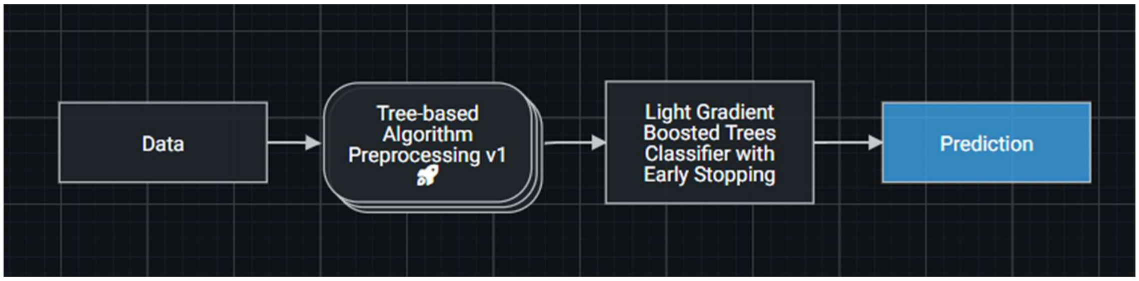

Data
Tree-based Algorithm Preprocessing v1
Light Gradient Boosted Trees Classifier with Early Stopping
Prediction

Fig 2.0 Light Gradient Boosted Trees Classifier with early stopping pipeline

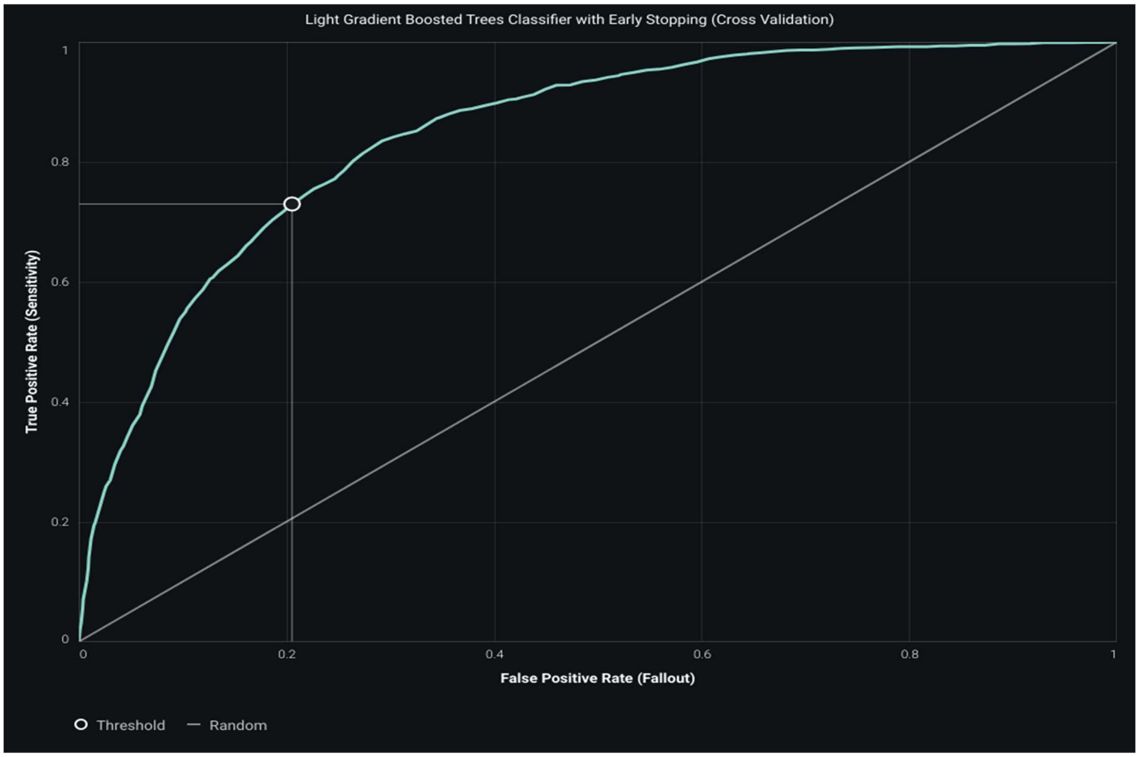


Fig 2.1 ROC CURVE(DATAROBOT)

Raw Data access: **:** [Telecom Churn Prediction | Kaggle](https://www.kaggle.com/code/bandiatindra/telecom-churn-prediction/data?select=WA_Fn-UseC_-Telco-Customer-Churn.csv) (https://www.kaggle.com/code/bandiatindra/telecom-churn-prediction/data?select=WA_Fn-UseC_-Telco-Customer-Churn.csv)

Referencing list

Docs.datarobot.com. 2022. *DataRobot Product Documentation*. [online] Available at: <https://docs.datarobot.com/>