**COMPARISON AND CONTRAST OF POINT & CLICK, AUTOML AND PROGRAMMING**

**ANALYTICAL TOOLS FOR PREDICTING CUSTOMER CHURN IN THE TELECOMMUNICATION INDUSTRY.**

**1.0 INTRODUCTION:**

Customer churn, also known as customer attrition, is one major problem facing businesses today. Research shows that winning new customers is up to ten times more expensive than retaining existing ones in today's competitive market. As a result, companies must focus on developing accurate and reliable predictive models for identifying customers who are likely to churn (Çelik & Osmanoglu, 2019).

Customer churn is a phenomenon that occurs especially among telecommunications service businesses. Recently, the telecommunications industry has undergone many changes, such as market liberalization, which opened the competition, and new technologies and services. Customer churn results in a huge loss of telecommunications services, which is a very serious problem (Huang, et al., 2012). With this in mind, the aim of this project is to use analytic tools such as Point & Click (RapidMiner and SAS Viya), AutoML (Datarobot) and Programming (Python) to predict customer churn which will help us know customers who are not going to renew their annual, monthly or daily subscription.

**DATASETS TO BE CONSIDERED**

* 1. **SAS VIYA (DATASET)**

The commsdata in SAS VIYA (RAW DATA ACCESS: commsdata.sas7bdat), which contains information about customer behaviour was used under this section. The commsdata data set contains 128 columns and 56,577 rows. Some of the variables in the dataset include demographic information, product usage and type, billing data, and customer service interaction. Additionally, there is a binary variable which indicates whether customers churned or not.

**RapidMiner, Datarobot and Python (DATASET)**

The telecom customer churn data from Kaggle will be analyzed and used in this section [ RAW DATA ACCESS: www.kaggle.com/datasets/blastchar/telco-customer-churn]. The dataset contains 7043 rows and 21 columns. The churn is set as the target variable.

**2.0 Point and click (SAS Viya)**

SAS Viya is a cloud-native AI, (point-and-click) analytic and data management platform that supports the analytics life cycle, enabling management team to turn data into insights. Businesses mostly use SAS Viya to explore and address complex analytical challenges of the present while effortlessly scaling for the future. In addition, SAS Viya offers a wide range of analytical tools, including machine learning, data mining, and statistical analysis. It supports multiple programming languages such as Python and R.

However, SAS Viya can be costly compared to other cloud-based platforms. Also, Customization and configuration can be complex and require a skilled administrator.

The model studio section of SAS Viya was used for this analysis. The Model Studio in SAS Viya is an integrated visual environment that provides a suite of analytic data mining tools to facilitate end-to-end data mining analysis. Upon creation of the project, the dataset was partitioned into test and validation data (holdout data). The test data was used to build the model while the validation data is used to optimize the complexity of the model and find the sweet spot between bias and variance and assess the performance of the model. The pipeline in Figure 1.0 (Appendix) shows the data passed through various nodes to get our model comparison and prediction. Firstly, the replacement node was used to replace outliers with a specific value. Secondly, the transformation node was used to apply numerical transformation to the input variables. Thirdly, new features were added to the text mining node by using the text variable verbatim in the commsdata data source. Fourthly, the variable selection node was used to reduce the number of inputs for modelling before comparing models.

**INSIGHTS FROM USING SAS Viya ON EXPLORING THE TELECOM CUSTOMER CHURN DATA**

The Gradient model was the champion model for this project. The model was chosen based on the KS (Youden) for the validate partition (0.59) 93.86% of the validate partition which was correctly classified using the Gradient Boosting model. Also, as shown in Fig. 1.2, the five most crucial variables are the following: Handset Age Group, Total Days Over Plan, Days Suspended Last 6M, Transformed MB of Data Use Month 6, and Total Late Payments Lifetime. Figure 1.3 also shows the champion model's ROC curve.

A picture containing bar chart

Description automatically generated

Fig 1.1Assessment of all models

Chart, bar chart

Description automatically generated

Fig 1.2 Most Important Variables for Gradient Boosting Model.

Chart, line chart

Description automatically generated

Fig 1.3 SAS Viya ROC Curve

**2.1 RAPIDMINER**

RapidMiner is a point-and-click data science platform just like SAS Viya that allows users to build data mining and machine learning models without requiring extensive programming knowledge. It has a visual interface that enables users to drag and drop data sources, algorithms, and tools onto a canvas and connect them in a workflow. Its point-and-click interface makes it accessible to users with various levels of technical expertise and is a great choice for those who want to quickly analyze and model their data without needing to write code. Furthermore, RapidMiner allows users to perform a wide range of data analysis tasks, such as data cleaning, preprocessing, visualization, and predictive modelling. The platform provides a vast library of built-in algorithms and tools for data mining, statistical analysis, and machine learning, making it easy to apply these methods to a variety of datasets.

However, RapidMiner can be memory-intensive, and users may require a high-performance computer to handle larger datasets as was the case in my second project last semester, where it took hours to run my model pipelines.

Find below the pipeline our sample data went through in the rapid miner studio interface. Diagram

Description automatically generated Diagram

Description automatically generated

Fig. 2 Rapidminer pipeline

**2.2 DATAROBOT (AUTO ML)**

Datarobot is a machine learning platform for automating, assuring, and accelerating predictive analytics, helping data scientists and analysts build and deploy accurate predictive models in a fraction of the time required by other solutions. Datarobot makes it easy to use and optimize the most valuable open-source modelling techniques from R, Python, Spark, VW, XGBoost and more. Furthermore, Datarobot allows us to build a robust predictive model without writing a single code, unlike the use of programming languages.

However, Datarobot can be quite expensive to run, especially for small businesses and individual users. The platform is most targeted towards big organizations with significant budgets for data analysis. Also, though Datarobot offers a wide variety of machine learning algorithms and techniques, it may not be as customizable as other platforms. Users who require more control over the machine learning process may find Datarobot automation limiting**.**

**INSIGHTS FROM USING DATAROBOT**

From our analysis, we obtained the tree-based variable importance chart whichshows the relative importance of each feature relative to the most important feature for predicting customers who are likely to churn. The tenure is the most important independent variable or relative importance of all key features making up the model. This is shown in the figure below.

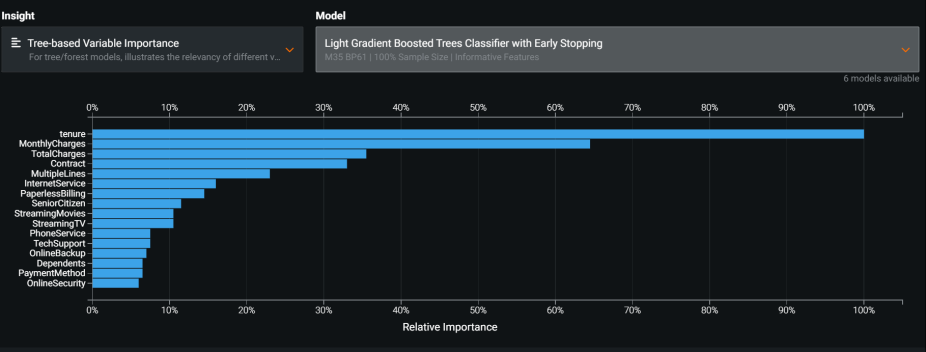


Fig 2.1 Feature importance chart of other variables to churn.

Also, the Light Gradient Boosted Tree classifier with early stopping is the champion model on the leaderboard.

Graphical user interface, diagram

Description automatically generated

Fig. 2.2 Blueprint for Light Gradient Boosted Trees Classifier with early stopping.

Line chart

Description automatically generated with low confidence Chart, line chart

Description automatically generated

Fig 2.3. Roc curve (Datarobot)

Fig 2.4. Lift chart (Datarobot)

Also, there is a model comparison tab in Datarobot to make comparisons among different models to know the different validation scores and prediction times.

A screenshot of a computer

Description automatically generated with medium confidence

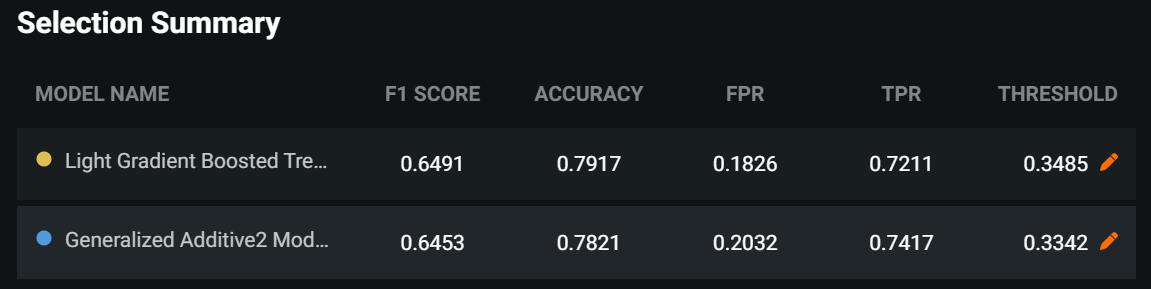


Fig 2.5. Model comparison and selection summary.

**3.0 PROGRAMMING (Python):**

Python is one of the most popular programming languages for data analytics. Also, it is a versatile language that can be used for various tasks such as data cleaning, visualization, machine learning, and lots more. Furthermore, Python's syntax is easy to learn and use, which can make it accessible to data analysts with varying levels of programming experience. With the help of Python libraries like pandas, summary statistics such as mean, standard deviation, minimum value, maximum value and percentile was easily obtained and this is done in less than a minute with the use of the pandas library “describe()” function in contrast to manual calculation In addition, using Python for data analysis, forecasting or prediction of data is simple to analyze using trends and graph plotting.

However, As Python continues to evolve, some libraries and tools may not be compatible with the latest version of the language, leading to compatibility issues and the need to maintain different versions of Python for different projects. Additionally, the vast ecosystem of libraries and tools available for Python can be overwhelming, and it may take some time to become proficient in selecting and using the most appropriate libraries for a given task.

**4.0 Conclusions**:

In predicting the best model for customer churn, SAS Viya and Rapidminer, AutoML (DataRobot) and Python were used. It was, however, handled differently to produce results. Datarobot is an automated machine-learning tool that is great for organizations with limited resources or expertise in data science. Additionally, Point & Click tools (such as SAS Viya and Rapidminer) need a data scientist to handle and adjust the parameters of these tools in order to achieve the best result. AutoML, on the other hand, is mostly automated, so even someone who isn't a data scientist or analyst can handle and tell the best prediction or classification model. Python is a flexible and widely used language that provides a lot of freedom and control over the entire data science workflow. However, Programming (Python) requires the services of a skilled programmer or data scientist who understands its functionalities in order to code and interpret outputs from the platform.

**Appendix**

Graphical user interface, diagram, application

Description automatically generated

Fig 1.0 Sas Viya pipeline

# References

Çelik, O. & Osmanoglu, U., 2019. Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments,* Volume 4(1), pp. pp.30-38.

Halawi, L., Clarke, A. & George, K., 2022. *Harnessing the Power of Analytics.* New York: Springer.

Huang, B., Kechadi, M. & Buckley, B., 2012. Customer churn prediction in telecommunications. *Expert Systems with Applications,* Volume 39(1), pp. pp.1414-1425.

https://www.coursera.org/learn/machine-learning-sas/home/info