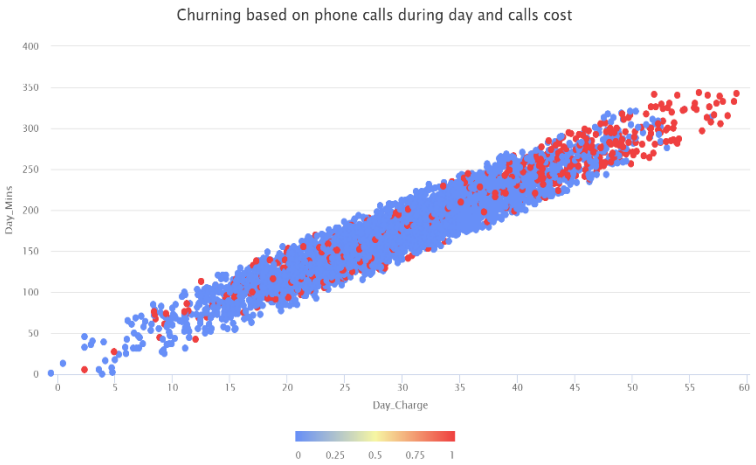
Project: A2 - Data Churn

Consultants: GHOSH Sounak – r0914328
SOLDATI Federico – r0924528

LAE
Leuven Analytics Experts & Partners

Part 1: Data exploration

Visual



Comments:

- The dataset presents no missing values.
- There is a linear relationship between the daily average of call minutes and daily average charges.
- There is a higher proportion of churning customers with a higher value of either daily minutes or daily charges.
- The visualization could suggest that, in proportion, customers who spend and use the service extensively are either those who save the most by switching operators, the less satisfied, or the most aware of offers/services from other providers.

Part 2: Modelling

2.a. Models

Technique 1: Gradient Boosting Trees (GBT)

Motivation:

- GBT can handle non-linear relationships between features and target variable
- GBT measures feature importance, helping grasp of contribution of variables.

Technique 2: k Nearest Neighbors (k-NN)

Motivation:

- Simple algorithm making it easy to understand and implement
- Outliers have less impact on k-NN

Technique 3: Random Forest (RF)

Motivation:

- RF provides high predictive accuracy.
- RF is robust to overfitting.

2.b. Evaluation

- For evaluation, measures like AUC-ROC and Accuracy were taken into account. Accuracy is very easy to understand, but it may give misleading results, particularly in the cases of imbalanced datasets. For this, along with Accuracy, AUC-ROC was implemented as it is a valuable tool in comparing the performance of various models making it a more appropriate metric in this case.
- To prevent overfitting, the 10-fold cross-validation was implemented on all the models.
- For GBT and RF, the maximum depth was set to 8 and 10 respectively, for k-NN the value of k was set to 5.
- The evaluation metrics are presented in the table in the results section, with higher values indicating better performance

Part 2: Modelling

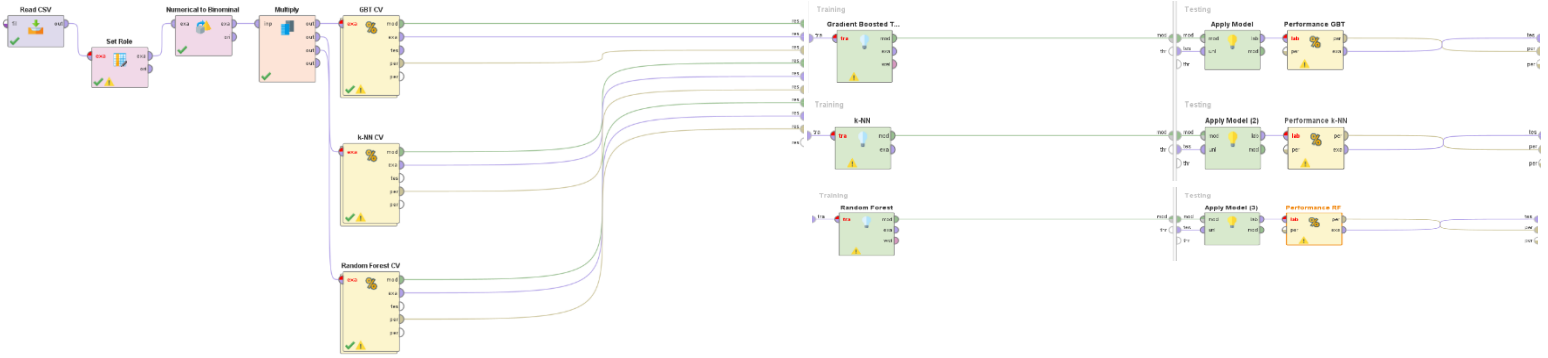
2.c. Results

Model	ACC	AUC
GBT	91.93% +/- 1.56%	0.871 +/- 0.029
K-NN	90.00% +/- 0.86%	0.732 +/- 0.040
RF	86.28% +/- 0.34%	0.734 +/- 0.060

Discussion:

- **Best performance:** Gradient Boosted Trees (GBT).
- Based on AUC, the best model is GBT followed by k-NN and then Random Forest. GBT has a far higher AUC score compared to the other models.
- According to GBT, the most important variables are Day_Mins, CustServ_Calls and Vmail_Message.
- Accuracy is also high for the GBT Model.

2.d. RapidMiner pipeline



Part 3: Pre-processing

1. Datasets often contain missing values and noisy data points. These data points, especially the noisy inconsistent ones, can influence the analysis greatly. Preprocessing the dataset involves the identification of such values which resolves all these issues, providing a great aid in the analysis.
2. Preprocessing allows data transformation like normalization, standardization and encoding of categorical variables. Many machine learning algorithms are very sensitive to the scale of the input features and require numerical input. Preprocessing addresses these challenges, consequently enhancing the functionality of machine learning models.
3. Preprocessing enables to create relevant features and reduce dimensionality in some cases, thus improving efficiency.