Churn Prediction Model

Context

- **This preliminary analysis** was made to show to a Telco Company the value of building an effective model for churn, using past transactional data and location data.
- The fitted model will be able to identify the **customers with most propensity to churn** along with suggestions of **possible further analysis.**

Objectives

- Present the descriptive analysis of the data.
- Describe the models used to solve the problem.
- Suggest the next steps.

Data preparation: scope & cleansing process

Target

- Develop a predictive model that using past data allow us to determine which customers has more probability to leave the company. Then, suggest further analysis.

Datasets and Variables to use

- One dataset with: Past transactional data and Location data for each customer

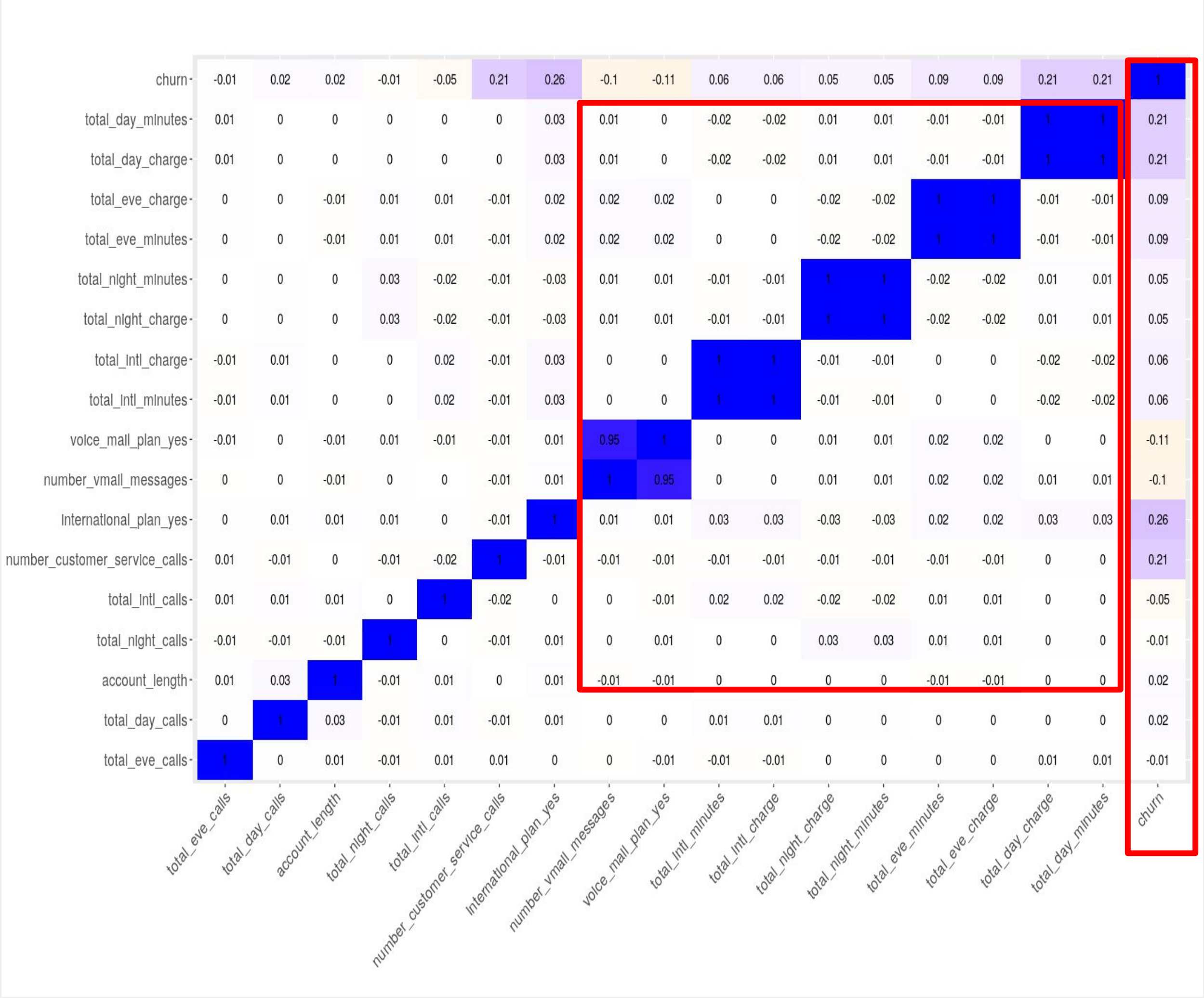
Cleansing process

- The acquired data has been prepared in the following way:

1. **Transform categorical/binary variables** (voicemail plan, international plan, state, area code, between others) **and numeric data** (create categorical variable for money spend in calls, between others).
2. **Filtering**: additional minor data filtering, … (here not was necessary)
3. **Cleaning inconsistencies:** Some variables were not considered in the model due to inconsistencies or due to poor predictive power or high correlations values (multicollinearity/noise).
4. **Creating new variables:** interactions between couple of variables can help to find more complex hidden patterns in the data that may improve the results (to explore in the future).

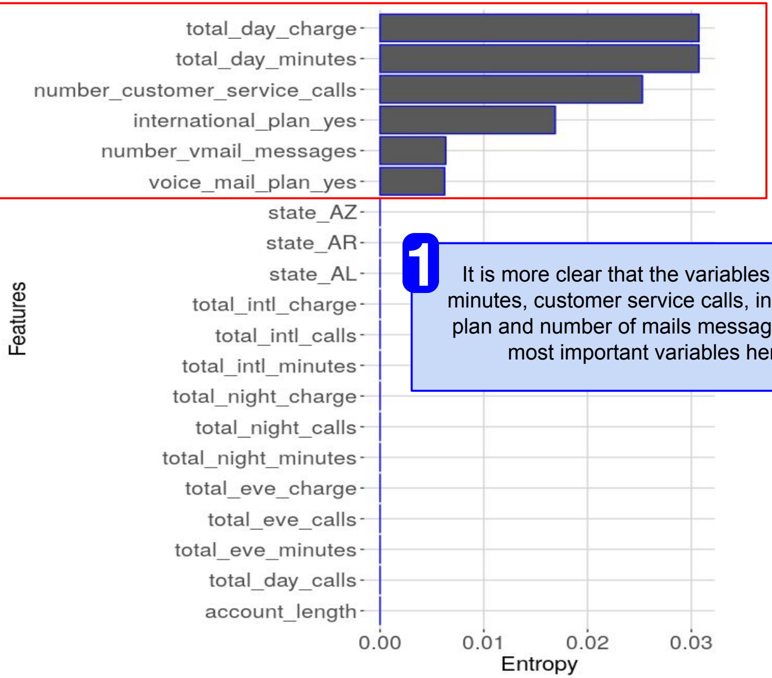
Correlations

- The variables more related with **Churn** are: international plan, number of customer service calls, total day charge, total day minutes, although these features are not highly correlated with churn
- Thera are variables that are highly correlated between them that can affect the results if they are all included. These are:
  - total_day_minutes vs total_day_charge
  - total_eve_minutes vs total_eve_charge
  - total_night_minutes vs total_night_charge
  - total_intl_minutes vs total_intl_charge
  - number vmail messages vs. voice_mail_plan
- We eliminate those variables that contains the variable charge. Although future analysis of consumptions would be required.
- We also eliminate the variable voice_mail_plan



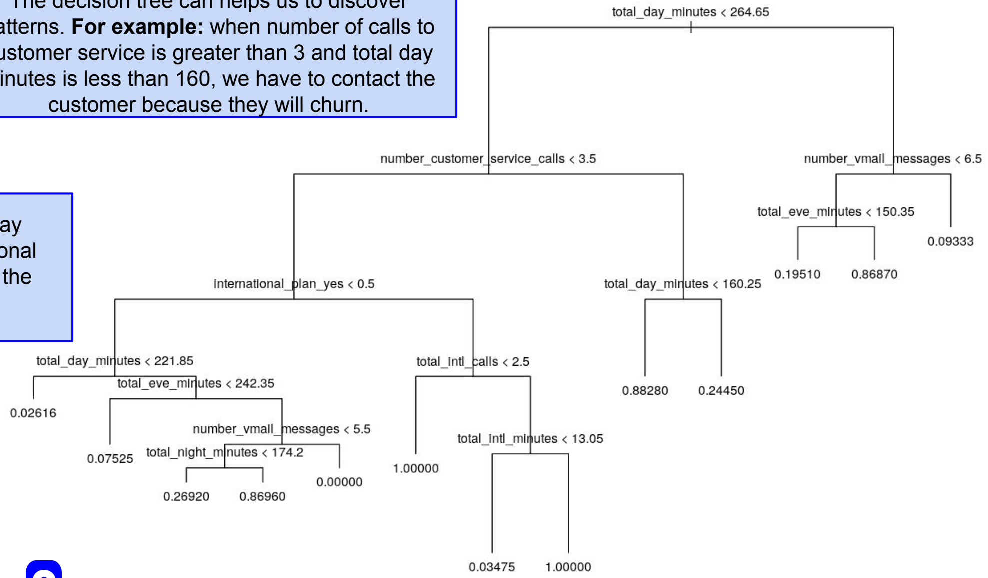
Dataset: Featuring Selection

- Using the **Entropy measure*** we can observe, which are the **variables that determine the propensity of Churn**. Additionally we can observe the **decision tree**, with the **rules of churn**



1 It is more clear that the variables total day minutes, customer service calls, international plan and number of mails messages are the most important variables here.**

1 The decision tree can helps us to discover patterns. **For example:** when number of calls to customer service is greater than 3 and total day minutes is less than 160, we have to contact the customer because they will churn.



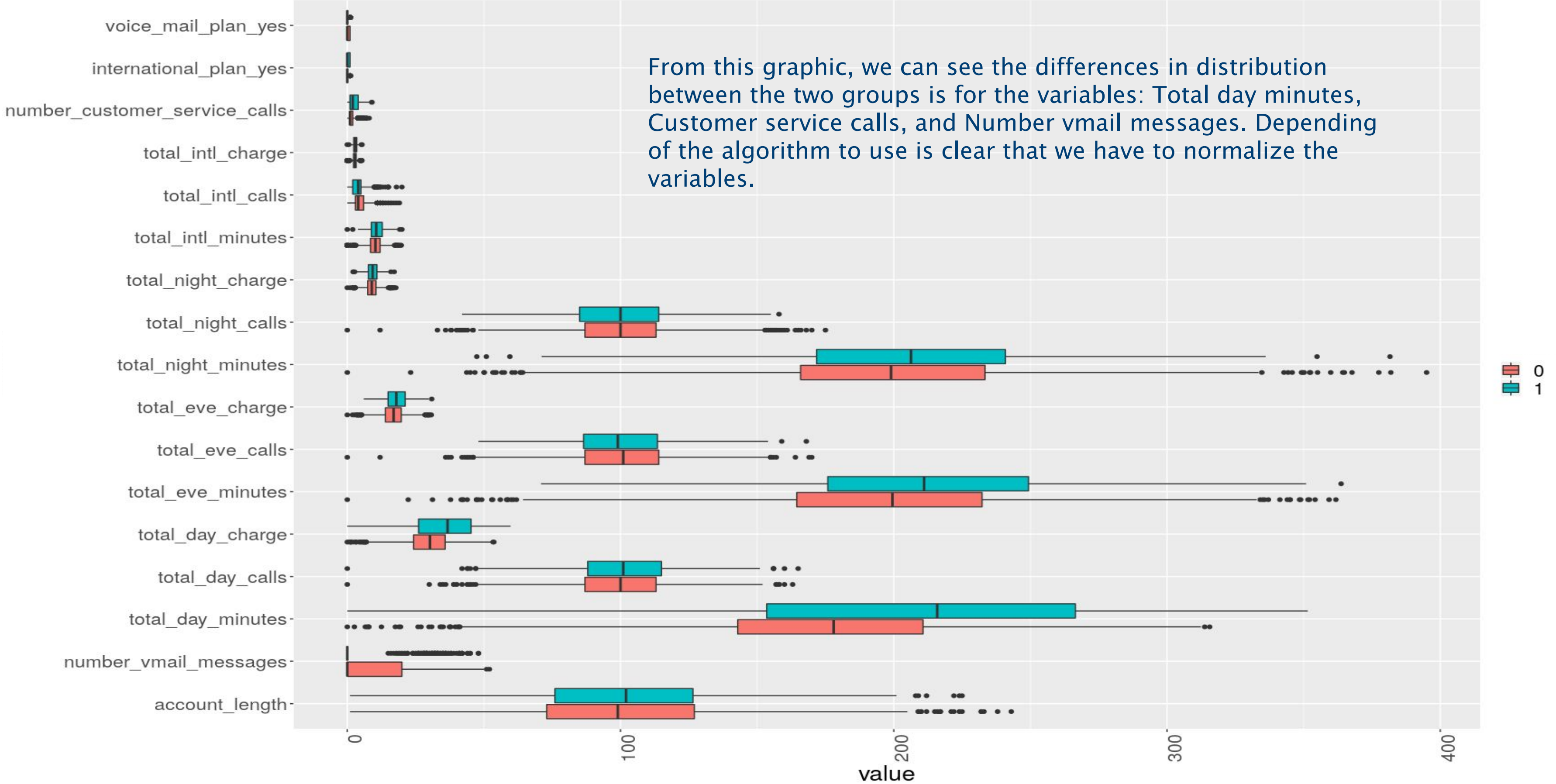
2 Investigate **interactions** between couple of variables can help to find more complex hidden patterns in the data that may improve the results. **Examples: customer service calls x International calls, customer service calls x total eve calls, etc.**

***Entropy measure** is used to calculate information gain that you obtain with each variable

**For the predictive models we will eliminate the variables related with charge and voice mail plan

Dataset: Featuring Selection

From this graphic, we can see the differences in distribution between the two groups is for the variables: Total day minutes, Customer service calls, and Number vmail messages. Depending of the algorithm to use is clear that we have to normalize the variables.



Dataset: Featuring Engineering

There are several ways to create new variables. Most of the where not considered in this analysis, but we describe the approaches for further work.

- Alternatives to explore:
  - Combine de variables related with charge and categorizes them in low, medium and high consumption.
  - Consider interactions between variables. Analyze which ones can bring additional value to the created model
  - Create additional variables that can bring more understanding about the drivers of the churn:
    - Call Diversity: Number of different persons called.
    - Cell Diversity: Did the client call from different locations?
    - Number of days since last call received (MTC -Mobile Termination Call)
    - Number of days since last call made (MOC -Mobile Originated Call)
    - Gap between calls
    - Promotions received
    - Mobile data consumed
    - Between others.

Applied methodology

The process to create a powerful and robust predictive model relies on **the following steps:**

Handling imbalanced data	There are several methods that help to remedy this problem. For this case, we applied **a technique called SMOTE*.**	No Churn aprox. 90% Churn aprox. 10%
Algorithms comparison & selection	We used **Logistic Regression (LR)**, **Gradient Boosted decision trees (GB)** and **Multilayer perceptron (MLP)** to predict the probability of a successful sale.	Value to minimize: **Classification Error**
Parameters optimization	We spent some time to optimize the parameters of the algorithms (using spark & python) in order to obtain the best results.	**5-fold cross val** LR:0.20 GB: 0.021 MLP: 0.11
Robust validation	First, we used **5-fold cross-validation** technique to train and test the models. Later, we **tested the model again** on an independent test set, not used for the training.	**Error** LR:0.24 GB: 0.054 MLP: 0.17

*SMOTE (Synthetic Minority Oversampling TEchnique) consists of add elements of the minority class, based on those that already exist. It works randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors.

Results

With a **test set of 30%** of the original data, the model seems to **validate the past pattern of the churn** (confusion matrix). Now, with the **second dataset** we will **estimate the value** of the model based in the performance of the theoretical model in the future (expected theoretical value).

Test Set (30%)

Confusion matrix (Threshold 0.5)

Random Forest (RF)

		Prediction		
		Loss	Win	TOTAL
Actual	No	991	307	1298
Value	SI	52	150	202
TOTAL		1043	457	1500

Accuracy	**0.76**
Precision	**0.33**
Recall	**0.74**
f1-score	**0.76**

Gradient Boosted Tree (GB)

		Prediction		
		Loss	Win	TOTAL
Actual	No	1258	40	1298
Value	Si	41	161	202
TOTAL		1299	201	1500

Accuracy	**0.95**
Precision	**0.80**
Recall	**0.80**
f1-score	**0.80**

Multilayer Perceptron

		Prediction		
		Loss	Win	TOTAL
Actual	No	1100	198	1298
Value	Si	58	144	202
TOTAL		1158	342	1500

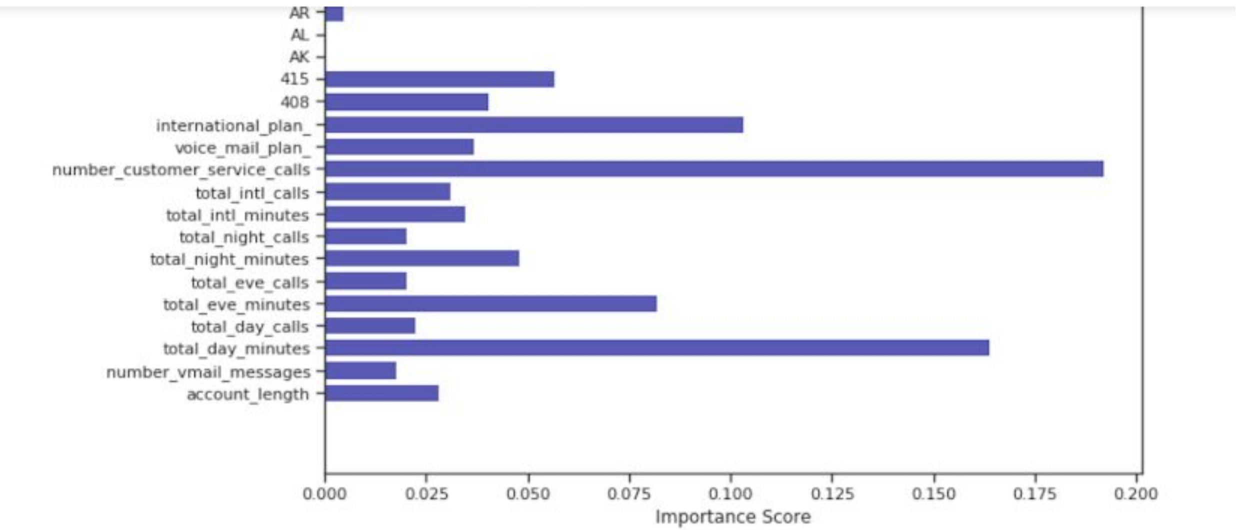
Accuracy	**0.83**
Precision	**0.42**
Recall	**0.71**
f1-score	**0.52**

Clearly, Gradient Boosted trees gives the best results. These results can be improved:

- Reducing the number of variables, to those that are more relevance (p-value analysis in the logistic regression, for example)
- Changing the parameters of the different models.
- Changing the thresholds for the probabilities of churn/no churn.
- Considering other methodologies.
- Investigate the possibility of include new variables

Featuring Importance

- Considering Xgboost as the best model, we can see that the variables that more contribute in the detection of churn are those that you can see in the figure. It is clear the number to customer service calls is a good proxi for churn, total day minutes and international plan seems to be very important as well.



Conclusions & Recommendations*

- We could observe that there is several variables that help to determine churn. A recommendation is to take care of those clients that make **calls to customer services**, also take care of **those that have changes in their consumption during the day** and also take care of **those clients that have international plan**.
- It is possible to **create rules** that help to specify promotions and measures to apply to those that seems to have more propensity to churn. These can be done through decision trees, association rules, between others.
- The **performance of the models could be improved** by incorporating additional features which include dynamic interactions (time), and more interactions between variables.
- In the future, it might be possible to **perform a controlled experiment** using the results of the machine learning algorithms to generate better understanding and data for simulations.
- This **work has been made using R + python and spark**. The combinations of these tools in one script can be used to get the best of each one.