**Hackathon Case Study**

*The following case competition is designed to test your ability to leverage advanced analytics/ machine learning to solve real world business problems and communicate your findings and derive insight from your technical work.*

*You may perform the analyses* ***using either R or Python as per your convenience.****. Please submit well commented codes along with detailed results. You may also use excel spreadsheet/ powerpoint to share detailed description of each of the steps performed along with outcome.*

**Business Scenario:**

Our client, ElectricCo, is a major utility company providing gas and electricity to **corporate, SME and residential customers**. In recent years, post-liberalization of the energy market in **Europe,** ElectricCo has had a growing problem with **increasing customer defections above industry average**. They would like **to identify the drivers of this problem and to devise and implement a strategy to counter it**. The churn issue is most acute in the SME division and thus they want it to be the first priority. The head of the SME division has asked whether it is possible to predict the customers which are most likely to churn so that they can trial a range of pre-emptive actions.

**Your task:**

The client also would like to answer the following questions:

1. What are the most explicative variables indicating churn?
2. Is there a correlation between subscribed power and consumption?
3. Is there a link between sales channel and churn?
4. Build a machine learning model that can predict the customers with high probability of churn.

The first stage is to establish the **viability** of such a model. For training your model you are provided with a dataset **which includes features of SME customers in January 2016** as well as the information about whether or not they have churned by March 2016. In addition to that you have received the prices from 2015 for these customers. **While it is not mandatory, but you are encouraged to test multiple algorithms to build predictive model.**

Using the trained model you shall “score” customers in the verification data set (provided in the eponymous file) and put them in descending order of the propensity to churn. You should also classify these customers into two classes: those which you predict to churn are to be labelled "1" and the remaining customers should be labelled "0" in the result template.

**Information contained in the data set**:

The below table describes all the data fields which are found in the data (across three files). You will notice that the contents of some fields are meaningless text strings. This is due to "hashing" of text fields for data privacy. While their commercial interpretation is lost as a result of the hashing, they may still have predictive power.

A whole host of rich investigations are possible. Your ideas on what some next steps could be, armed with such data is also of interest.

Data fields and their description

There are 16096 SME customers present. 15500 customers considering for Model training, validation and testing and 596 customers considered for predicting going to churn in next 3 months or not (where output is not present for these 596 customers)

Table: Hackthon\_case\_training\_data – The data provided for all the 16096 customers

|  |  |  |
| --- | --- | --- |
| **Field name** | **Description** | **Present** |
| id | contact id |  |
| activity\_new | category of the company's activity | 6551 |
| campaign\_disc\_ele | code of the electricity campaign the customer last subscribed to (none) |  |
| channel\_sales | code of the sales channel | 11878 |
| cons\_12m | electricity consumption of the past 12 months |  |
| cons\_gas\_12m | gas consumption of the past 12 months |  |
| cons\_last\_month | electricity consumption of the last month |  |
| date\_activ | date of activation of the contract |  |
| date\_end | registered date of the end of the contract |  |
| date\_first\_activ | date of first contract of the client | 3508 |
| date\_modif\_prod | date of last modification of the product |  |
| date\_renewal | date of the next contract renewal |  |
| forecast\_base\_bill\_ele | forecasted electricity bill baseline for next month | 3508 |
| forecast\_base\_bill\_year | forecasted electricity bill baseline for calendar year | 3508 |
| forecast\_bill\_12m | forecasted electricity bill baseline for 12 months | 3508 |
| forecast\_cons | forecasted electricity consumption for next month | 3508 |
| forecast\_cons\_12m | forecasted electricity consumption for next 12 months |  |
| forecast\_cons\_year | forecasted electricity consumption for next calendar year |  |
| forecast\_discount\_energy | forecasted value of current discount |  |
| forecast\_meter\_rent\_12m | forecasted bill of meter rental for the next 12 months |  |
| forecast\_price\_energy\_p1 | forecasted energy price for 1st period |  |
| forecast\_price\_energy\_p2 | forecasted energy price for 2nd period |  |
| forecast\_price\_pow\_p1 | forecasted power price for 1st period |  |
| has\_gas | indicated if client is also a gas client |  |
| imp\_cons | current paid consumption |  |
| margin\_gross\_pow\_ele | gross margin on power subscription |  |
| margin\_net\_pow\_ele | net margin on power subscription |  |
| nb\_prod\_act | number of active products and services |  |
| net\_margin | total net margin |  |
| num\_years\_antig | antiquity of the client (in number of years) |  |
| origin\_up | code of the electricity campaign the customer first subscribed to |  |
| pow\_max | subscribed power |  |
| price\_date | reference date |  |

Table2: Hackthon\_case\_training\_hist\_data – The data provided for all the 16096 customers – Every customer have last 12 months data (12 rows of data for each Customer )

|  |  |  |
| --- | --- | --- |
| id | contact id |  |
| price\_p1\_var | price of energy for the 1st period |  |
| price\_p2\_var | price of energy for the 2nd period |  |
| price\_p3\_var | price of **energy** for the 3rd period |  |
| price\_p1\_fix | price of **power** for the 1st period |  |
| price\_p2\_fix | price of power for the 2nd period |  |
| price\_p3\_fix | price of power for the 3rd period |  |

Table3: Hackthon\_case\_training\_output – The output provided for only 15500 customers

|  |  |  |
| --- | --- | --- |
| id | contact id |  |
| churned | has the client churned over the next 3 months |  |

Table4: Sample\_output – The output required provided for only 596 customers (required to provide probability and predicted value output)