

CAPSTONE REPORT

Predicting Customer Churn at QWE.Inc

**Summary of Findings**

Data Science Consulting

Ingrid Ortega Rojas

2020

# Introduction

This report is about a data science analysis in QWE Inc., a dot-com start-up which helps small and medium size businesses manage their online presence through a subscription service. The company is in the mature growing phase therefore customer retention results critical. The desired output from this analysis are insights that guide the responsible of the department to take preventive actions in customers which are more probable of leaving.

The data science process used was BADIR:

A screenshot of a cell phone

Description automatically generated

The programming language to analyze data and develop the models is Python. Code was written and provided in a Jupyter Notebook which is available in GibHub:

<https://github.com/ingridortega/module5>

# 1. Business questions

**General question:** Is it possible to determine the probability a customer leave within the next 2 months in order to take preventive actions?

**Background:** Customer service actually at QWE has a reactive approach when a customer tries to leave. There are 3 types of contracts: month, 6 months or 12 months. The VP of the department is the stakeholder who is wondering if the probability that a customer leaves within 2 months can be determined in order to enhance accordingly his or her experience.

**Business considerations:** The analysis should be done a short timeline less than 2 weeks, therefore the VP can apply the improvement to present to the CEO of QWE who is the sponsor for the project. Another stakeholder is a VP associate which is interested in the success of the project as he extracted the data.

# 2. Analysis plan

**Main goal:** Develop model to predict customers that are going to churn in QWE within the next 2 months.

**Specific goals:**

* Identify factors that affect customer churn.
* Identify characteristics of customers.

**Type of analysis:** supervised learning, clasification (label = default 1/0).Test different supervised algorithms (eg. Logistic, SVM, trees) and choose one that gives more confidence to answer the business question.

**Hypothesis to be tested:**

* “…if a customer has been with us for more than 14 months, he or she knows how to use our services, is using them, gets values of it, and should be less likely to leave.”
* “Those between 6 and 14 months are probably the riskiest group”.
* “Those with high CHI don´t leave much”

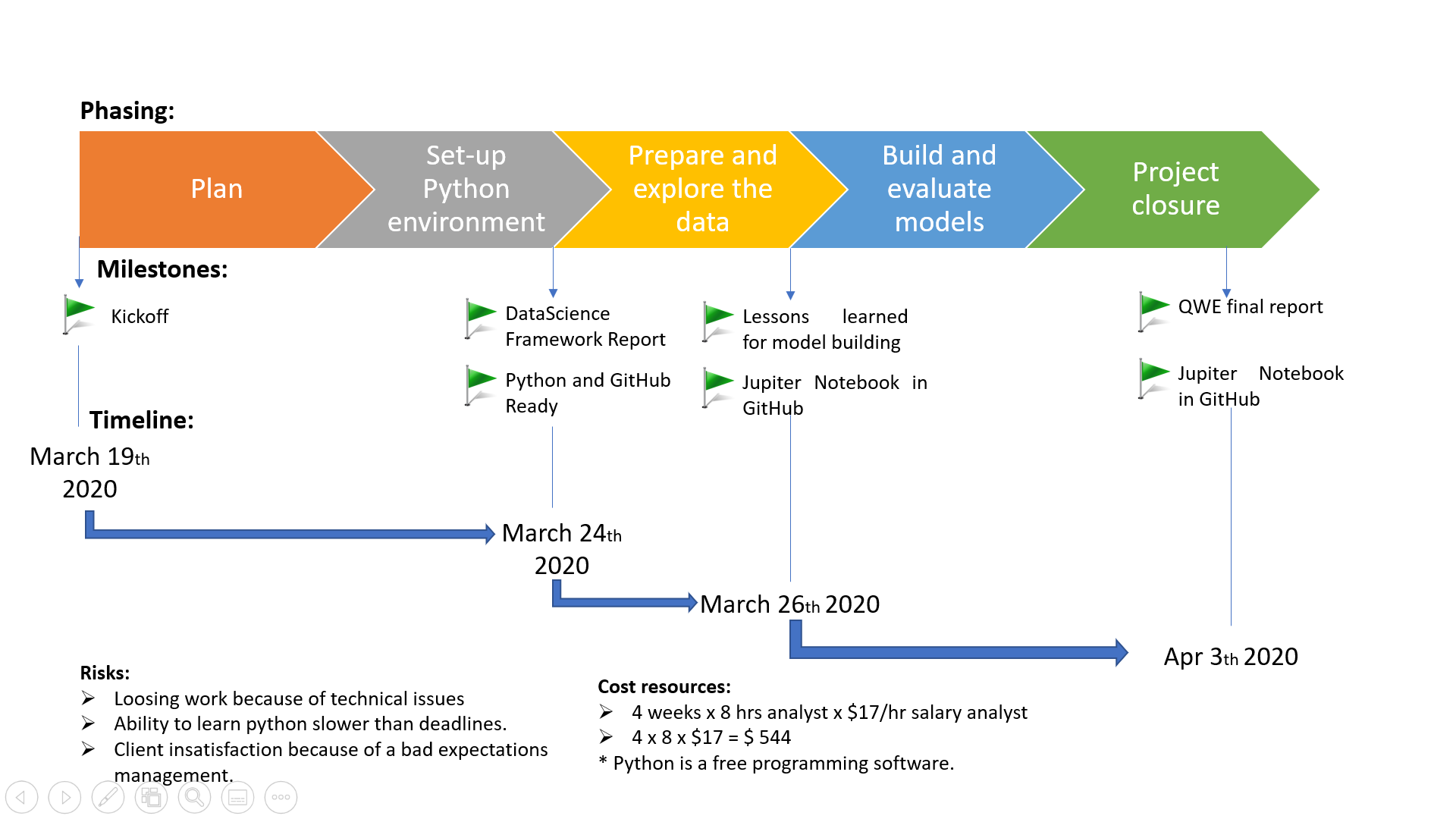
**Data to test hypothesis:** Sample of QWE’s customers obtained from salesforce.com database rolled back 2 months to December 1st 2011. **Structure of data:** 6.000 registries, 13 items (column), 1 identifier, 12 attributes, 1 dependent variable.

* **Customer age (months):** Customer longevity
* **Churn (1=yes, 0=no):** Dependent variable
* **CHI month 0:** Customer Happiness Index of current month (December 2011)\*
* **CHI 0-1:** CHI December 2011 – CHI November 2011
* **Support Cases month 0:** Number of support cases current month (December 2011)
* **Support Cases 0-1:** Support Cases December 2011 – Support Cases November 2011
* **SP month 0:** Average support priority current month (December 2011)
* **SP 0-1:** Average support priority month before current (November 2011)
* **Logins 0-1:** Usage information: Logins December 2011 – Logins November 2011
* **Blog Articles 0-1:** Usage information: Blogs December 2011 – Blogs November 2011
* **Views 0-1:** Usage information: Views December 2011 – Views November 2011
* **Days since last login 0-1:** Days since last login December 2011 – Days since last login November 2011

\*CHI combines multiple characteristics related to customers experiences with services.

**Methodology to be employed:** train, test, validate and compare performance metrics (accuracy, precision, recall) of different classification supervised algorithms: eg.logistic regression, decision trees, KNN, SVM, Naive Bayes.

**Project plan:**



# 3. Data collection, preprocessing and feature engineering

**Data source:** the data corresponds to a Harvard Business Publishing Education case.

* Case description:

<https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&ved=2ahUKEwjchN-2kKDoAhUB7qwKHerDCIkQFjADegQIBRAC&url=https%3A%2F%2Fclemson.instructure.com%2Ffiles%2F2114955%2Fdownload%3Fdownload_frd%3D1%26verifier%3DoHu4GeAXapg2MKMZmz8rqARFcJQn4zeLDEBkSqKb&usg=AOvVaw36cRcGQ8MdXVN45kpo0BAA>

* Supplementary material from case downloaded from url:

<https://hbsp.harvard.edu/product/UV6694-PDF-ENG>

**Cleanse and validation of data:**

* Data was revised looking for missing /null values which were going to be either replaced or ignored.
* Validation: after cleansing; missing values will be looked to see if they remain.

# 4. Insight from data

* 1. Answers to investigation questions
* Yes, it is possible to determine churn with 60% of probability, by running the model.
* “…if a customer has been with us for more than 14 months, he or she knows how to use our services, is using them, gets values of it, and should be less likely to leave.”

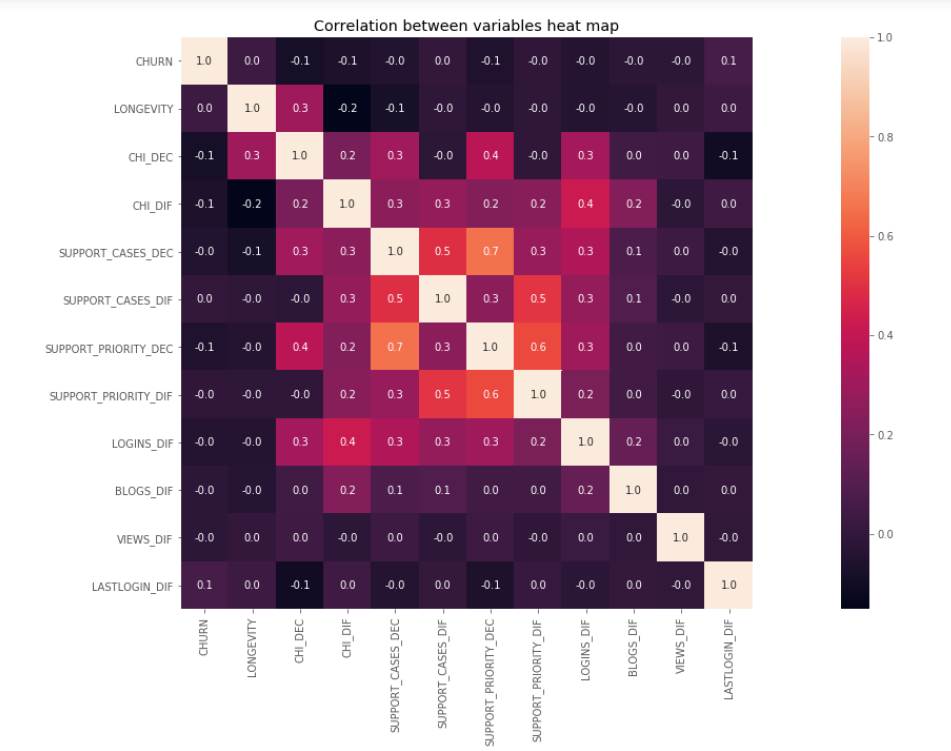
The data showed after if for example using a logistic approach, that the less likely to leave is the group between 7 and 13.

* “Those between 6 and 14 months are probably the riskiest group”.

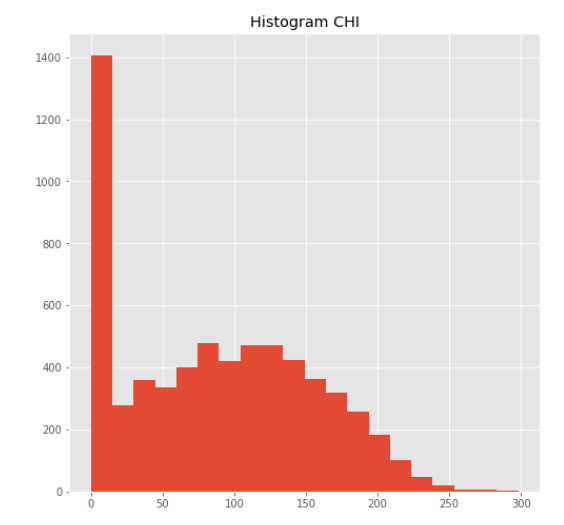
No, after running the model are the least risky.

* 1. Attributes in the data statistically significant to the problem

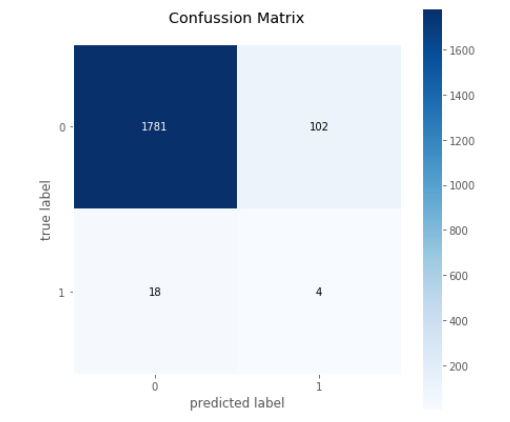
The last login was the most important attribute.

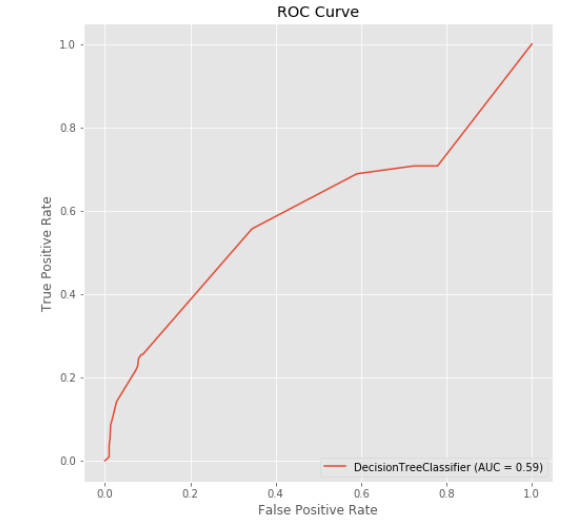


* 1. Relevant visualizations





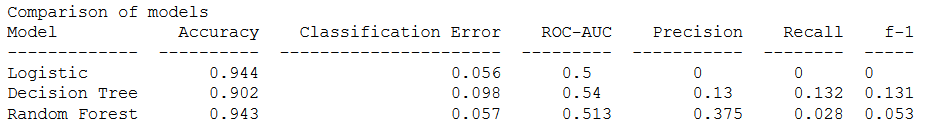




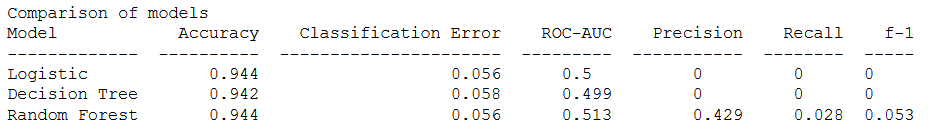
* 1. Other observations

The following are the results from the rounds of train-test:

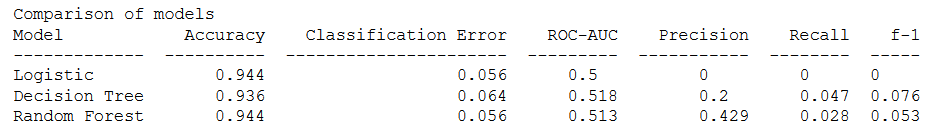
Round 1: (here is where the decision tree classifier was selected, as it’s recall is the higher; the others were overfitting).



Round 2:



Round 3:

+

* 1. Rules that provide insights

**Model selection:** churn can be predicted with a decision tree classifier.

# 5. Final recommendations

* Evaluate all the insights and take actions that improve actual performance.
* Automatize this model, including it in the software for client service.
* Update the model yearly according to the new behavior of data.