**INTRODUCTION**

Using data to make informed decisions is the best decision. Every day the amount of data available grows exponentially .as a result, effective interpretation is more important than ever. Data analytics is quickly becoming one of the world’s most exciting and rewarding career paths. Business analytics skills will most likely be in higher demand over the next decade than any other career (10.9% vs 5.2%). career of labor statistics).

Companies all over the world require qualified data analysts to solve problems and assist them in making the best business decisions possible. Currently ,59% of companies intend to add even more positions requiring data analysis skills (source: SHRM).

In this article, I will describe the process in achieving my outcome. This is a classification project. this project, is all imaginary, supposing every company wants to increase its profits or revenue margin and customers retention is one key area industry players want to focus on. In today’s world of machine learning, most companies build classification models to perform churn analysis on their customers.

Classification in machine learning and statistics is a supervised learning approach in which the computer program learns from the data given to it and makes new observations or classifications.

The main goal is to identify which class the new data will fall into.

This project will allow us learn more about classification models and help the clients, a telecommunication company, to understand their data. Breathing life and meaning to data in its raw form.

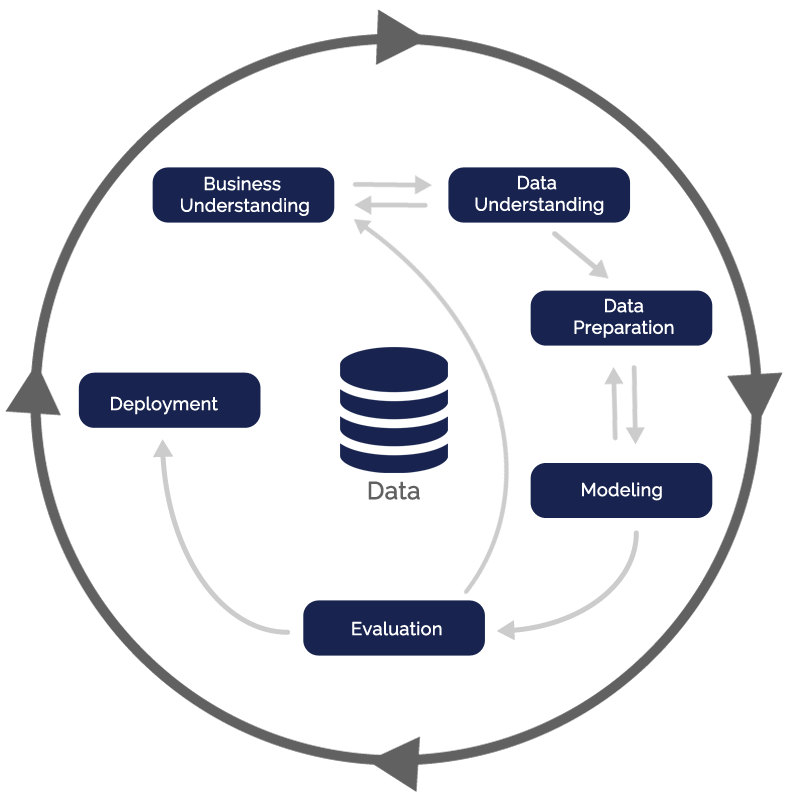
Also, to find the lifetime value of each customer.

Furthermore, know the factors that affect the rate at which customers stop using their network.

Lastly, to predict if a cluster will churn or not.

In the end, I hope to have explored all options on the data, modelled development using machine learning algorithms like logistic regression, decision trees, support vector machine, random forest etc., I hope to have developed a model evaluation and interpretation using LIME, SHP techniques, model optimization and hyperparameter tuning. I hope to have asked the right questions, summarized data, connected business objectives to data analysis, identified and cleaned the data, created visualizations, and above all I hope to have told a data driven story.

Every data science or data analytics project follows a certain kind of data science process. Scrum, Kanban, and Agile are all methods that data science teams adopt to complete their projects but, in this project, I will be working with CRISP-DM.

I used the cross industry standard process for data mining (CRISP-DM) model as the base for my data science process.it has six sequential phases:

1. Business understanding - what does the business need?
2. Data understanding – what data do we have/ need? Is it clean?
3. Data preparation – how do I organize the data for modelling?
4. Modelling- what modelling techniques should I apply?
5. Evaluation – which model best meets the business objectives?
6. Deployment- how do stakeholders access the results?

**THE DATA**

This project aims to identify the key indicators of customer churn for a telecommunications company and develop a model to predict which customers are likely to churn. The project will also provide insights into effective retention strategies that the company can implement to reduce customer churn. The data will be processed and analyzed using various techniques such as data cleaning, bivariate and multivariate analysis, and exploratory data analysis. The best-performing model will be selected and evaluated, and suggestions for model improvement will be provided. The ultimate goal of this project is to help the telecommunications company reduce customer churn and improve customer retention.

**File Descriptions and Data Field Information**

Telco customer.csv

* The training data, contain about 21 columns

**ASK STAGE**

At this stage, we bring the objective into view and put down the questions that we intend to answer at the end of the analysis process. The first phrase of the data analysis process is asking the right questions.

Here, with the overarching goal of making a recommendation to the team to assist their goal of entering the Indian startup ecosystem, the whole picture was considered to make sure we get the situation right. The following hypothesis was stated and questions were asked to guide the analyses.

it's important to use new data when evaluating our model to prevent the likelihood of overfitting to the training set. However, sometimes it's useful to evaluate our model as we're building it to find that best parameters of a model - but we can't use the test set for this evaluation or else we'll end up selecting the parameters that perform best on the test data but maybe not the parameters that generalize best.

**HYPOTHESIS**

Five clearly stated null and alternate hypotheses are:

1. H0: There is no significant difference in churn rates between male and female customers.
2. H1: There is a significant difference in churn rates between male and female customers.
3. H0: There is no significant relationship between the customer's internet service provider and their likelihood to churn.
4. H1: There is a significant relationship between the customer's internet service provider and their likelihood to churn.
5. H0: There is no significant difference in churn rates between customers who have paperless billing and those who don't.
6. H1: There is a significant difference in churn rates between customers who have paperless billing and those who don't.
7. H0: There is no significant difference in churn rates between customers with a senior citizen status and those without.
8. H1: There is a significant difference in churn rates between customers with a senior citizen status and those without.
9. H0: There is no significant difference in churn rates between customers on different types of payment methods.
10. H1: There is a significant difference in churn rates between customers on different types of payment methods.

**RESEARCH QUESTIONS**

1. What percentage of customers have churned?
2. Is there a correlation between a customer's length of tenure with the company and their likelihood of churning?
3. Are there any specific groups of customers based on demographic that are more likely to churn than others?
4. Can customer retention be improved by offering longer contract terms?
5. How much money could the company save by reducing customer churn?
6. What is the relationship between Internet Services and churn rate?
7. Are senior citizens likely to churn more?

**DATA PREPARATION AND PROCESSING**

Here I organize the data to make it fit for analysis. Cleanliness and consistency of data are the objectives promotion? to make sure that all datatypes are correct.

Here are a few steps that you can use to validate your time series machine learning models:

* Compare the results of your model with those of a baseline method, such as a simple moving average.
* Compare the predictions of your model against actual data.
* Use rolling windows to test how well the model performs on data that is one step or several steps ahead of the current time point.
* Compare the predictions of your model against those made by a human expert.
* Use machine learning techniques, such as k-fold cross-validation, to test the generalization accuracy of your model.

**LOADING PACKAGES**

To start with, the basic packages for analysis were loaded into my jupyter notebook. These packages were:

Pandas: for data cleaning and manipulation

NumPy: for data cleaning and manipulation

Glob: a module that has several functions, that can help in listing files under a specified folder.

Matplotlib: visualization tool

Seaborn

# Library for EDA

import pandas as pd

import NumPy as np

import seaborn as suns

%Matplotlib inline

import matplotlib. pilot as plt

import matplotlib. dates as mdates

from sklearn. impute import SimpleImputer

from pandas\_profiling import ProfileReport

import warnings

warnings. filterwarnings('ignore')

**GENERAL NOTES FROM PREVIEWING THE DATA FRAMES**

* All the columns with amounts have to be set to float.
* Upon examining the data frame, I discovered that some of the columns contain values that should be numerical but are currently strings (objects). I will need to convert the datatypes of the values in these columns to numerical (float and/or integer).
* Also, there were a considerable number of null values in the datasets
* Data inspected for null values

**ASSUMPTIONS**

1. Imputations will not be made for undisclosed and/or unavailable (missing) amounts due to the uncertainties, risks of misstatements and possible misleading effects on the analyses.
2. All other things been equal (ceteris paribus)

**DATA CLEANING**

the major activities performed on the Data Frames with respect to data cleaning are explain below.

The detailed functions will be found in the jupyter notebook, a link to which will be attached at the end of the article.

**ANSWERING RESEARCH QUESTIONS**

At this point, I combine the analyses and share stages of the data analysis process the coding and visualization of the merged data.

**PLEASE CLICK ON THIS LINK TO ACCESS THE JUPYTER NOTEBOOK:** **https://github.com/arkularyea/classification-project.git**

**CONCLUSION**

TO SEE THE CONCLUSION PLEASE CLICK ON THE LINK TO ACCESS THE JUPYTER NOTEBOOK: https://github.com/arkularyea/classification-project.git

**FINAL NOTES**

Thank you so much for reading.

I will be grateful for your comments, advice, suggestions, and recommendations. Too long or too short? Too detailed or missing some details? Please let me know. You can leave a comment here or find me on Twitter (@arku\_laryea).

GitHub link: https://github.com/arkularyea/classification-project.git

Medium link: https://medium.com/@nlaryea70/project-classification-ef5597e0f1f

LinkedIn link: https://www.linkedin.com/pulse/project-classification-nii-laryea