Capstone Two: Project Ideas & Proposal

# Idea 1: Customer Churn prediction

## Dataset:

Telco customer churn: IBM dataset

<https://community.ibm.com/accelerators/catalog/content/Telco-customer-churn>

## **Context**

A fictional telco company that provided home phone and Internet services to 7043 customers in California in Q3.

## **Data Description**

7043 observations with 33 variables

## Project Scope:

Review the customer data and get insights about customer churn, try to see any trends of customers leaving the company and the possible reason behind it. Also build a machine learning model to predict probability of customer leaving the company

# Idea 2: Investment prediction

## Project Scope:

Forecast future value of stock based on historical data and current trends.

## Dataset:

Web scraping of data from yahoo finance.

https://ca.finance.yahoo.com/

# Idea 3: Predict purchasing capabilities of customer (customer segmentation)

## Dataset:

<https://www.kaggle.com/frtgnn/dunnhumby-the-complete-journey?select=product.csv>

This dataset contains household level transactions over two years from a group of 2,500 households who are frequent shoppers at a retailer. It contains all of each household’s purchases, not just those from a limited number of categories. For certain households, demographic information as well as direct marketing contact history are included.

## Project Scope:

Review the customer data and make predictions on purchasing capabilities of a customer.

# Idea 4: Predict the estimated risk of a patient's chance of getting thyroid disease

## Dataset:

<https://archive.ics.uci.edu/ml/datasets/thyroid+disease>

## Data Set Information:

# From Garavan Institute  
# Documentation: as given by Ross Quinlan  
# 6 databases from the Garavan Institute in Sydney, Australia  
# Approximately the following for each database:  
  
\*\* 2800 training (data) instances and 972 test instances  
\*\* Plenty of missing data  
\*\* 29 or so attributes, either Boolean or continuously-valued  
  
# 2 additional databases, also from Ross Quinlan, are also here  
  
\*\* Hypothyroid.data and sick-euthyroid.data  
\*\* Quinlan believes that these databases have been corrupted  
\*\* Their format is highly similar to the other databases  
  
# 1 more database of 9172 instances that cover 20 classes, and a related domain theory  
# Another thyroid database from Stefan Aeberhard  
  
\*\* 3 classes, 215 instances, 5 attributes  
\*\* No missing values  
  
# A Thyroid database suited for training ANNs  
  
\*\* 3 classes  
\*\* 3772 training instances, 3428 testing instances  
\*\* Includes cost data (donated by Peter Turney)

## Project Scope:

Review the available data and build a machine learning model for predicting a patient’s risk level.