National College of Ireland

MSc Data Analytics

Data Mining Machine Learning - I

Project Pre Proposal

Viplav Gadewar

Student ID – x21164274

Department of Computing

Lecturer

Dr. Pierpaolo Dondio

Department of Computing

06 March 2022

Research Questions

1 - How precise can we make a system that predicts an apartment's rental price?

2 - How well can we anticipate customer satisfaction with airline service?

3 - How accurately can we estimate the amount of impurities in the iron ore concentrate? The percentage of silica present can be used to determine impurity.

Data sets

Data Set 1 – "Apartment rental offers in Germany"

Rental offers scraped from Germany biggest real estate online platform. The data was scraped from https://www.immobilienscout24.de/, the biggest real estate platform in Germany. immobilienscout24 has listings for both rental properties and homes for sale, however, the data only contains offers for rental properties.

Source: https://www.immobilienscout24.de/

Dimensions: 268,850 rows and 49 columns

Out of 49 columns only 17 columns will be used for rental price prediction to get precise results.

Below are column definitions for 17 columns:

- regio1: federal state of Germany.
- heatingType: Type of heating.
- newlyConst: is the building newly constructed?
- balcony: does the object have a balcony?
- yearConstructed: year of construction
- hasKitchen: has a kitchen?
- cellar: has a cellar
- livingSpace: living space in sqm
- condition: condition of the flat
- interiorQual: interior quality
- lift: is elevator available?
- typeOfFlat: type of flat
- geo_plz: ZIP code
- noRooms: number of rooms
- floor: which floor is the flat on
- garden: has a garden?
- baseRent: rent of the apartment.

Target Variable: baseRent

Data Set 2 – “Twitter US Airline Sentiment”

Analyse how travellers in February 2015 expressed their feelings on Twitter. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service").

Source: https://www.kaggle.com/

Dimensions: 14,640 rows and 15 columns

Out of 15 columns only 3 columns are used for doing sentiment analysis of airline service.

Below are column definitions for 3 columns:

- airline_sentiment: positive, negative, or neutral
- text: text present in tweet
- airline: name of the airline

Target Variable: airline_sentiment

Data Set 3 – “Quality Prediction in a Mining Process”

Explore real industrial data and help manufacturing plants to be more efficient. This dataset contains data from range (March of 2017 until September of 2017). Some columns were sampled every 20 second. Others were sampled on an hourly base.

Source: https://www.kaggle.com/

Dimensions: 737,453 rows and 24 columns

Out of 24 columns only 12 columns are taken to predict quality of iron ore based on silica present.

Below are column definitions for 12 columns:

- Date: Date collection date and time.
- % Iron Feed: Feed grade of iron-containing ore.
- % Silica Feed: Feed grade of silica-containing ore.
- Starch Flow: Depressant chemical for Iron (Fe) containing ore.
- Amina Flow: Collector chemical for Silica containing ore.
- Ore Pulp Flow: The amount of pulp flow fed to the Flotation Columns as the product of the previous process step.
- Ore Pulp pH: pH.
- Ore Pulp Density: The solid percent of ore fed to Flotation Columns.
- Flotation Column 01,02,03,04,05,06,07 Air Flow: The amount of air fed to the Flotations Columns to frothing.
- Flotation Column 01,02,03,04,05,06,07 Level: Showing float thickness of Flotation Columns.
- % Iron Concentrate: Concentrate grade of iron-containing ore.
- % Silica Concentrate: Concentrate grade of silica-containing ore

Target Variable: % Silica Concentrate