M.Sc. Data Analytics and Technologies


# DAT7303- Data Mining and Machine Learning

## Assessment 1- Portfolio 3

Submitted By:

Tahmain Akther Monisa

Student ID: 2310413

Submitted To:

Pradeep Hewage

Module Instructor

Date: 26th April, 2024

Time of workshop session: 13:00 pm

**Table of contents**

## 1.0 Introduction

This report contains tasks related to analysing real-world datasets, including data preprocessing, feature selection, model selection, performance evaluation, and interpretation of results, and demonstrates the ability to deal with complex issues systematically. Each task follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology framework to ensure a systematic and structured approach to the analysis process. The CRISP-DM methodology is a widely-used and has become the most common methodology for data mining, analytics, and data science projects. CRISP-DM consists of six main phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

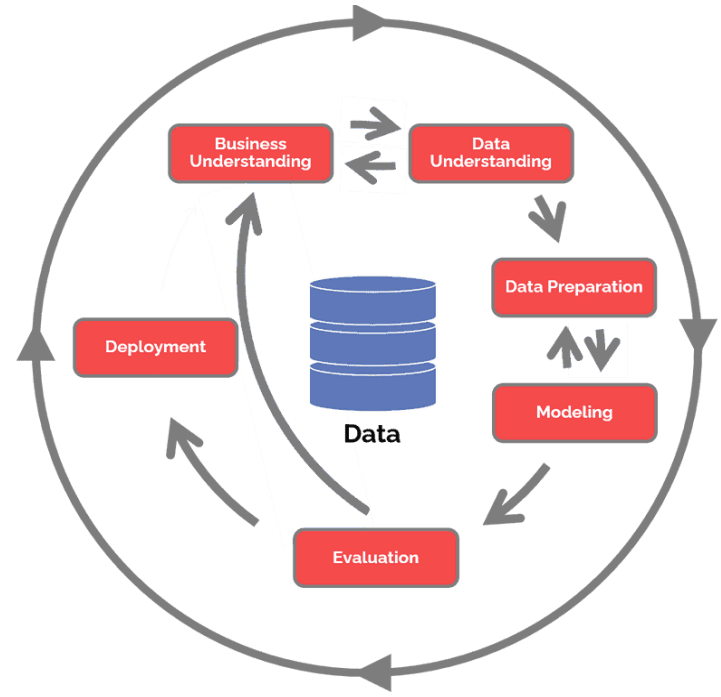


Figure 1: CRISP-DM Diagram.

“These six phases with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict. In fact, most projects move back and forth between phases as necessary” (IBM

Documentation, 2021). The phases are not in a fixed order. It is constantly necessary to switch back and forth between phases. Which phase, or specific task within a phase, needs to be completed next is determined by the results of each phase. The most significant and common interphase dependencies are shown by the arrows. For this report, Dataset named “miami-housing” is used. From importing data, cleaning, preparing data, analysing data, appropriate predictive analytical techniques, appropriate visualisations, and analysis results has been applied. Accurate housing price forecasts are essential for many stakeholders in today's dynamic real estate market, including buyers, sellers, and real estate professionals. By adopting data-driven approaches, one can gain important insights into the variables affecting property prices, which can help with strategic planning and well-informed decision-making.

## 2.0 Business Understanding

“The first phase of the CRISP-DM methodology focuses on understanding the project's objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives” (Wirth and Hipp, 2000).

### 2.1 Determine business objectives

To ensure that the data mining project is in line with the demands and objectives of real estate market stakeholders, it is imperative to have a thorough understanding of the business objectives. For the Miami Housing Dataset, Accurate home price estimation is necessary to assist sellers' pricing tactics and buyers' purchase decisions. Determining the key elements of house pricing in order to guide marketing and investment choices. enhancing customer satisfaction and market transparency by offering trustworthy insights into the Miami real estate market. For success criteria, Attain low root mean squared error (RMSE) and mean absolute error (MAE) while projecting house prices. Provide stakeholders with useful information that they may use to make better decisions and achieve better business results.

### 2.2 Assess the situation

This task entails gathering more precise information on all the resources, limitations, presumptions, and other elements that need to be taken into account while establishing the project strategy and goal for the data analysis. Identify available

data sources, including the Miami Housing Dataset and any additional data that may be required for analysis.

### 2.3 Determine data mining goals

Creating predictive models that determine the price of houses depending on features like size, location, and facilities. Determining the main determinants of housing prices, such as market trends, neighbourhood demographics, and economic indicators. To effectively target certain market segments, divide the Miami housing market into submarkets based on buyer preferences and property attributes.

### 2.4 Produce project plan

Choosing appropriate data mining tools and technologies, such as R or SQL, for data analysis and modelling. Considering additional tools for data visualisation, feature engineering, and model deployment as needed.

## 3.0 Data Understanding

At this point, the data is examined in further detail. Because it offers comprehensive information about the selected dataset, this phase is critical. There are several ways to understand data.

### 3.1 Collect initial data

The first step is to obtain the required information from reliable sources. Regarding the Miami Housing Dataset, the information could come from government organisations, real estate databases, or other sources. After acquiring it, the data must be entered into analysis programs like Python, R, or SQL to be processed further. The data set chosen for this project is the Miami Housing dataset.

### 3.2 Describe data

There are 17 columns and 13933 rows in the dataset. This dataset offers detailed information on residential properties in the Miami region, including geographic coordinates, physical attributes, sale prices, and environmental elements like proximity to water, highways, and rail lines. The dataset contains the following columns:

| **Attributes** | **Description** |
|---|---|
| PARCELNO | unique identifier for each property. About 1% appear multiple times. |
| SALE_PRC | sale price ($) |
| LND_SQFOOT | land area (square feet) |
| TOTLVGAREA | floor area (square feet) |
| SPECFEATVAL | value of special features (e.g., swimming pools) ($) |
| RAIL_DIST | distance to the nearest rail line (an indicator of noise) (feet) |
| OCEAN_DIST | distance to the ocean (feet) |
| WATER_DIST: | distance to the nearest body of water (feet) |
| CNTR_DIST | distance to the Miami central business district |

| | |
|---|---|
| SUBCNTR_DI | distance to the nearest subcenter (feet) |
| HWY_DIST | distance to the nearest highway (an indicator of noise) (feet) |
| age | age of the structure |
| avno60plus | dummy variable for aeroplane noise exceeding an acceptable level |
| structure_quality | quality of the structure |
| month_sold | sale month in 2016 (1 = jan) |
| LATITUDE | LATITUDE |
| LONGITUDE | LONGITUDE |

Table 1: Attributes Description of the Dataset

### 3.3 Explore data quality

Utilise methods such as exploratory data analysis (visualising trends and relationships) and data profiling (summarising important characteristics) to gain a high-level knowledge of the data.

### 3.4 Verify data quality

Check the data for problems with quality, such as outliers, inconsistencies, and missing numbers. This guarantees the accuracy of the data for modelling.

## 4.0 Data preparation

This phase involves cleaning and preparing the data for modelling. This may involve handling missing values, correcting inconsistencies, transforming data into a suitable format, and feature engineering (creating new features from existing ones).

## 4.1 Data cleaning

Take care of previously found problems in data quality. Imputing missing numbers, fixing discrepancies, and managing outliers may all be part of this. For “miami housing” dataset, longitude, latitude and avno60plus columns are replaced to NULL.

## 4.2 Data transformation

Convert data into a modelling-ready format. This could entail feature engineering (making new features), encoding categorical variables, or scaling numerical features.

## 4.3 Data integration

To ensure consistency and reduce redundancy, combine data from several sources into a single, cohesive dataset.

## 5.0 Modelling

In this phase, modelling techniques are selected and applied. This could involve various algorithms like decision trees, regression analysis, or clustering, depending on the specific problem needed to solve.

### 5.1 Select modelling techniques

Determining which algorithms to try. For this particular dataset regression algorithms are used to create models. For instance, Support vector regression, Decision Tree, and Random forest.

### 5.2 Generate test design

Train the chosen model on a portion of the prepared data. This involves splitting the data into training and testing sets. For this scenario, 70% of the data is being used for training and the rest is for testing.

### 5.3 Build model

Following selection, a variety of algorithms and techniques are used to train the models on the training data. In order to reduce error and increase predicted accuracy, model parameters are optimised during the training process.

### 5.4 Assess model

The models' performance is assessed using the testing data after training. Depending on the kind of problem being handled, this evaluation entails analysing measures like accuracy, precision, recall, or mean squared error.

## 6.0 Evaluation

During the evaluation phase, the produced models' performance is assessed about the project's goals and success criteria. In this phase, model performance metrics are analysed, the findings are interpreted, and recommendations are based on the analysis. Based on the analysis and interpretation of the results, The best-fitting model is the random forest model. Low Root mean square error (RMSE) score indicates the Random forest n500 model is the best model performance.

## 7.0 Deployment

The derived models are put into production systems for practical application during the deployment phase. This stage includes training stakeholders on how to use the models, integrating them into current business processes, and tracking their effectiveness over time. The first step in model deployment is incorporating the created models into stakeholder applications or production systems. The models can be accessed and used to enhance decision-making and drive business outcomes thanks to this integration.

## 8.0 Conclusion

This data mining project attempts to offer important insights into Miami housing price prediction by utilising the CRISP-DM approach. Stakeholders will be able to make well-informed decisions in the Miami real estate market by comprehending company objectives, examining and organising the data, creating predictive models, assessing performance, and implementing actionable insights. In the fast-paced real estate sector, the initiative helps to improve client satisfaction, decision-making processes, and market transparency.

**References**


1. IBM Documentation. (2021). SPSS Modeler Subscription. [online]. Available from: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview [Accessed April 26, 2024].
2. Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. [online]. Available from: https://www.semanticscholar.org/paper/CRISP-DM%3A-Towards-a-Standard-Process-Model-for-Data-Wirth-Hipp/48b9293cfd4297f855867ca278f7069abc6a9c24?utm_source=topic_pages_prototype [Accessed March 8, 2024].