

Course/Programme:	MSc. Artificial Intelligence
Module Name and Code:	Data Mining and Machine Learning (DAT7303)
Student ID:	2011184
Student Name:	Bogomil Iliev
Tutor:	Dr. Pradeep Hewage
Assessment Number:	3 of 4
Assessment Type:	Portfolio
Assessment Title:	Portfolio 3
Indicative Word Count:	At least 2000 words.
Weighting:	30% of overall module grade.
Submission Deadline:	18.04.2025 (no later than 23:59)
Submission Date:	18.04.2025 at 23:00
Learning Outcomes assessed: LO1: Critically review the principles, theories, algorithms and techniques used in Data Mining and Machine Learning then apply them creatively to solve complex problems. LO2: Critically apply data mining and machine learning techniques to real-world datasets, including data preprocessing, feature selection, model selection, performance evaluation, and interpretation of results, and demonstrate the ability to deal with complex issues systematically.	

Data-Driven Valuation of Residential Properties in Bolton.

Abstract

This study applies the CRISP-DM methodology to construct and evaluate predictive models for residential property prices in the Bolton region. A dataset of 13 932 samples, containing structural attributes, locational distances and environmental indicators, was cleaned, transformed and analysed entirely in R. Irrelevant geographical coordinates and unique identification were removed. Six algorithms have been trained and tested to outline the best performing one that is going to be the best fit for the business purposes. The Random Forest model achieved the best generalisation accuracy, outperforming the rest. The model managed to forecast an accurate valuation of \$ 1.45m on a pre-defined abode requirements by the stakeholders.

Keywords: Random Forest, House Price Prediction, CRISP-DM, Bolton.

Contents

Abstract	iii
Table of Figures.....	vi
Table of Tables.....	vi
1. Introduction.....	1
2. Business Understanding.....	2
2.1. Business Objectives.....	2
2.1.1. Business Problem Definition.	2
2.1.2. Success Criteria.....	2
2.2. Evaluation of Situation.....	3
2.2.1. Resource Availability.....	3
2.2.2. Risk Assessment.....	3
2.2.3. Cost-Benefit Analysis.....	4
2.3. Data Mining Goals.....	5
2.4. Project Plan.....	5
3. Data Understanding Phase.....	6
3.1. Initial Data Collection.....	6
3.2. Data Description (Preliminary Exploration).....	6
3.3. Data Exploration (Preliminary Analysis).....	12
3.4. Data Quality Verification.....	14
4. Data Preparation Phase (Data Munging).....	17
4.1. Selecting the data.....	17
4.2. Data Cleaning.....	17
4.3. Construct the Data (Feature Engineering).....	18
4.4. Integration of Data (Formatting for Modelling).....	18
5. Modelling Phase.....	19

5.1.	Selection of Model Techniques.	19
5.1.1.	Multiple Linear Regression (MLR).	19
5.1.2.	Support-Vector Regression (SVR).	19
5.1.3.	Decision Tree.	20
5.1.4.	Random Forest.	20
5.2.	Testing Design.	21
5.3.	Model Building.	21
5.4.	Assessing of Model Performance.	22
6.	Evaluation Phase.	24
6.1.	Evaluation of Results Against Business Goals.	24
6.2.	Review of Process.	24
6.3.	Prediction Generation for a Specified Property.	25
6.4.	Next Steps.	25
6.4.1.	Deployment Plan.	25
6.4.2.	Monitoring and Maintenance.	25
6.4.3.	Future Improvements.	25
7.	Conclusion.	26
8.	Bibliography.	27
9.	Word Count.	29
10.	GAI Declaration.	29
11.	Appendices.	30
11.1.	List of Abbreviations Used.	30

Table of Figures

Figure 1 - First Ten Rows from the Bolton Housing Prices Dataset.....	6
Figure 2 - Last Ten Rows from the Bolton Housing Prices Dataset.	7
Figure 3 - Provided Information Regarding the Columns of the Dataset by the Stakeholders.	8
Figure 4 - Dataset Structure.	10
Figure 5 - Dataset Summary.....	10
Figure 6 - Detailed Dataset Summary.....	11
Figure 7 - Histogram for Distribution of Sale Prices.	12
Figure 8 - Box Plot of House Sale Prices.....	13
Figure 9 - Correlational Heatmap of Housing Features.	14
Figure 10 - Missing Values per Column Check.....	15
Figure 11 - Checking for out-of-range values in the "age" and "structure_quality" columns.	15
Figure 12 - Multiple Occurances of the Same Parcel Number in the Dataset.....	16
Figure 13 - Checking for Potential Data Anomalies.	16
Figure 14 - RMSE Comparison Result Bar Chart.	23

Table of Tables

Table 1 - Cost-Benefits Table.	4
Table 2 - Initial Project Plan.....	5
Table 3 - Data Summary Analysis and Insights Table	12
Table 4 - Test Set Evaluation Result Table.	23
Table 5 - Review of Process Table.....	24

1. Introduction

The expansion in employment of data-driven solutions throughout the last decade in the decision-making process across various industries has notably escalated the value of accurate predictive analytics. In the real estate sector, dependable forecasting of property sale prices allows key players to improve their financial commitment capabilities in regard to the investments choices they make. It also alleviates their perception of possible abode costs fluctuations and planning, thus enhancing their ability to remain competitive on the market. Hence, the current project focuses on applying the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology in order to deliver dependable predictive models to accurately foretell residential asset costs in the Bolton area. Through the provided house prices dataset from 2023 in Bolton, that gives information regarding property specific features such as land and living area, proximity to amenities, structure quality, and noise indicators, the current work systematically covers the steps of business and data understanding, data preparation, modelling, and evaluation (Putler and Krider, 2012; Larose and Larose, 2019; Hamizah Zulkifley *et al.*, 2020).

Pivotal for this procedure is the thoughtful and justified choice of models that are capable of performing the latter task, comprehensive analysis of the dataset, and rigorous assessment of model performance against the outlined business and technical success criteria. The project emphasises on delivering high predictive accuracy and interpretability via thorough data preprocessing techniques, insightful visualisations, and in-depth evaluations. Thus, guaranteeing the derived knowledge is practically applicable to the business (James *et al.*, 2021).

2. Business Understanding.

This phase of the project allows for a better comprehension of the patron's business necessities, and it highlights a clear outline of what is targeted and needed by performing such a project (Putler and Krider, 2012).

2.1. Business Objectives.

2.1.1. Business Problem Definition.

The central business necessity around which the current work revolves is to **foretell abode sale prices in the Bolton residential area accurately**. In such a manner stakeholders' (real estate agents, homebuyers, and investors) abilities to make informed financial decisions is uplifted. Precise cost projections amplify market competitiveness, better cost planning, and apprise asset commitments (Hamizah Zulkifley *et al.*, 2020; Larose and Larose, 2019).

2.1.2. Success Criteria.

In order to guarantee the balance betwixt the technical aspects of data mining and the business value, it is pivotal to delineate an explicit success criterion for the project. In such a way, the insights delivered by it would hold worth for the business (Putler and Krider, 2012). Hence, the following are defined as **project success criteria** for the current work (Putler and Krider, 2012; James *et al.*, 2021):

- **Accuracy** – Shrinking error metrics.
- **Interpretability** – The solution must deliver comprehensible data revelations into key determinants that guide property costs.
- **Generalisability** - The proposed solution should be robust and foretell dependably even on new information.

2.2. Evaluation of Situation.

In order for a Data Mining (DM) project to be fruitful, it should take into consideration the potential barriers that might occur during its development and what the business is anticipating as returns (Witten *et al.*, 2017).

2.2.1. Resource Availability.

To perform the current project the following resources have been highlighted as important:

- **Bolton Housing Prices Dataset** – the information has been provided by the company alongside a detailed explanation of different property features.
- **Software & Tools** - R programming environment for analysis, visualisation, and modelling purposes (including the respective libraries for the tasks).
- **Skills and Knowledge Required** – Proficiency in data preprocessing techniques, regression analysis, and machine learning (ML) modelling in an R environment.

2.2.2. Risk Assessment.

The possible risks that are associated with the aforementioned project, which may need mitigation are (Fawcett and Provost, 2013):

- **Quality of Information** – absent or conflicting data entries may diminish the accuracy delivered by the models.
- **Model Interpretability** – Sophisticated architectures (such as deep learning ones) might lessen the trust stakeholders put in the project due to the unclear decision-making process involved into their development.
- **Overfitting** – the solution must achieve similar results on both training and unfamiliar information.

2.2.3. Cost-Benefit Analysis.

Administering a cost-benefit analysis is of utmost importance to decide on the economic viability and strategic justification of a data mining project. Calculating the spending involved in resource management, hardware needed, time and skill-levels, against the expected returns of a project deliver a better way of addressing the decision of whether a project should be conducted or discontinued. Accordingly, stakeholders can make more informed decisions on which activities should be prioritised and where more resources should be spent (Shmueli *et al.*, 2017; Fawcett and Provost, 2013).

Hence, in **Table 1** are outlined some of the major Costs and Benefits involved in the current project.

Table 1 - Cost-Benefits Table.

Benefits	Improved valuation accuracy
	Competative advantage
	Better market insighths
Costs	Time spent in data preprocessing
	Time spent in data modelling
	Computational resources
	Training costs for model users

2.3. Data Mining Goals.

The understandable outlining of technical objectives warrants that the models developed in a data mining project will definitely achieve the business goals and can be used within its production and operations (Han *et al.*, 2011).

Thus, in technical terms, the primary objectives that are identified are:

- **Identification of key factors that govern housing prices.**
- **Deliver predictive models that can proficiently foretell sale prices.**
- **Maintain equilibrium among prediction accuracy and model transparency.**

2.4. Project Plan.

For a project to be fruitful, planning of its stages, tasks, responsibilities and tools needed is pivotal. In such a manner is guaranteed a beneficial outcome of its delivery and the dangers of stagnation points is diminished. Hence, in **Table 2** is drafted the proposed initial project plan (Putler and Krider, 2012).

Table 2 - Initial Project Plan.

CRISP-DM Phase	Activity	Tool (R)	Duration
Business & Data Understanding	Initial Exploration & Documentation	dplyr, ggplot2, summarytools	1-2 days
Data Preparation	Cleaning, Formatting, Feature Creation	tidyr, caret, mice, dplyr	2-3 days
Modeling	Model Development (3 methods minimum)	caret, randomForest, e1071	2-3 days
Evaluation	Model Performance Evaluation	caret, Metrics	1 day
Report Preparation	Comprehensive Reporting and Visuals	RMarkdown, MS Word	2 days

3. Data Understanding Phase.

This project phase is topping up on the insights gather from the previous one by placing the emphasis on locating and acquiring the needed information in a data frame and assessing the specific one that is going to be useful in achieving the project objectives (Putler and Krider, 2012).

3.1. Initial Data Collection.

Since the stakeholder organisation provided the Bolton Housing Prices dataset from 2023. There is no need to gather data externally or via web scraping.

3.2. Data Description (Preliminary Exploration).

In **Figures 1 & 2** (where the first and last ten records are portrayed) can be observed that the columns available in the data frame do coincide with the initial data feature information provided by the stakeholders (**Figure 3**).

	LATITUDE <dbl>	LONGITUDE <dbl>	PARCELNO <dbl>	SALE_PRC <dbl>	LND_SQFOOT <int>	TOT_LVG_AREA <int>	SPEC_FEAT_VAL <int>	RAIL_DIST <dbl>	OCEAN_DIST <dbl>	
1	25.89103	-80.16056	622280070620	440000	9375	1753	0	2815.9	12811.4	
2	25.89132	-80.15397	622280100460	349000	9375	1715	0	4359.1	10648.4	
3	25.89133	-80.15374	622280100470	800000	9375	2276	49206	4412.9	10574.1	
4	25.89176	-80.15266	622280100530	988000	12450	2058	10033	4585.0	10156.5	
5	25.89182	-80.15464	622280100200	755000	12800	1684	16681	4063.4	10836.8	
6	25.89206	-80.16135	622280070180	630000	9900	1531	2978	2391.4	13017.0	
7	25.89247	-80.15722	622280080100	1020000	10387	1753	23116	3277.4	11667.8	
8	25.89302	-80.15743	622280080400	850000	10272	1663	34933	3112.4	11718.1	
9	25.89305	-80.16156	622280080020	250000	9375	1493	11668	2081.8	13043.8	
10	25.89305	-80.15805	622280080370	1220000	13803	3077	34580	2937.7	11917.7	
⚡	RAIL_DIST <dbl>	OCEAN_DIST <dbl>	WATER_DIST <dbl>	CNTR_DIST <dbl>	SUBCNTR_DI <dbl>	HWY_DIST <dbl>	age <int>	avno60plus <int>	month_sold <int>	structure_quality <int>
	2815.9	12811.4	347.6	42815.3	37742.2	15954.9	67	0	8	4
	4359.1	10648.4	337.8	43504.9	37340.5	18125.0	63	0	9	4
	4412.9	10574.1	297.1	43530.4	37328.7	18200.5	61	0	2	4
	4585.0	10156.5	0.0	43797.5	37423.2	18514.4	63	0	9	4
	4063.4	10836.8	326.6	43599.7	37550.8	17903.4	42	0	7	4
	2391.4	13017.0	188.9	43135.1	38176.2	15687.2	41	0	2	4
	3277.4	11667.8	0.0	43598.7	37973.9	17068.2	63	0	2	5
	3112.4	11718.1	10.5	43780.8	38198.3	16989.9	21	0	9	4
	2081.8	13043.8	51.5	43481.7	38542.0	15623.3	56	0	3	4
	2937.7	11917.7	9.7	43730.1	38235.2	16787.0	63	0	11	5

1-10 of 10 rows | 9-18 of 17 columns

Figure 1 - First Ten Rows from the Bolton Housing Prices Dataset.

	LATITUDE <dbl>	LONGITUDE <dbl>	PARCELNO <dbl>	SALE_PRC <dbl>	LND_SQFO... <int>	TOT_LVG_AREA <int>	SPEC_FEAT_VAL <int>	RAIL_DIST <dbl>	OCEAN_DIST <dbl>	
13923	25.61086	-80.38105	3.050310e+12	245000	8000	1731	13686	6742.1	23833.3	
13924	25.61193	-80.38111	3.050310e+12	265000	8000	1346	7944	6629.0	23890.1	
13925	25.61217	-80.38285	3.059360e+12	230000	7500	1539	3474	6051.2	24470.3	
13926	25.78046	-80.26073	1.313201e+11	315000	9062	2261	2856	4221.7	19831.4	
13927	25.78280	-80.26076	1.313200e+11	215000	9605	1640	6856	3665.3	20593.9	
13928	25.78313	-80.25979	1.313200e+11	275000	6780	967	6580	3844.5	20568.0	
13929	25.78359	-80.26035	1.313200e+11	340000	7500	1854	2544	3593.6	20791.9	
13930	25.78379	-80.25613	1.313200e+11	287500	8460	1271	2064	4143.2	20307.9	
13931	25.78401	-80.25754	1.313200e+11	315000	7500	1613	3136	3986.9	20542.6	
13932	25.78439	-80.25890	1.313200e+11	250000	8833	1867	266	3793.9	20859.6	
	RAIL_DIST <dbl>	OCEAN_DIST <dbl>	WATER_DIST <dbl>	CNTR_DIST <dbl>	SUBCNTR_DI <dbl>	HWY_DIST <dbl>	age <int>	avno60plus <int>	month_sold <int>	structure_quality <int>
	6742.1	23833.3	10397.0	86423.6	35873.5	905.4	33	0	7	2
	6629.0	23890.1	10669.9	86168.4	35582.8	908.8	13	0	9	2
	6051.2	24470.3	11159.2	86523.5	35879.1	1478.7	33	0	5	2
	4221.7	19831.4	3030.7	22394.5	11189.9	1864.3	68	0	2	4
	3665.3	20593.9	2918.8	22467.2	12042.3	1022.3	69	0	3	2
	3844.5	20568.0	3252.4	22175.9	12150.1	917.4	16	0	4	4
	3593.6	20791.9	3077.7	22375.1	12316.8	738.2	26	0	5	4
	4143.2	20307.9	3588.4	20966.9	12433.0	743.7	16	0	7	4
	3986.9	20542.6	3589.1	21475.6	12458.0	626.1	16	0	8	4
	3793.9	20859.6	3421.0	21928.6	12599.0	474.7	62	0	11	4

1-10 of 10 rows | 9-18 of 17 columns

Figure 2 - Last Ten Rows from the Bolton Housing Prices Dataset.

The dataset contains the following columns:

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 central business district (feet)

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 airplane noise exceeding an acceptable level

structure_quality: quality of the structure

month_sold: sale month in 2023 (1 = jan)

LATITUDE

LONGITUDE

Figure 3 - Provided Information Regarding the Columns of the Dataset by the Stakeholders.

Figure 4 clearly shows that the number of records (rows) present in the dataset are 13 932, and the information is spread into 17 columns. Namely:

- LATITUDE – coordinates.
- LONGITUDE - coordinates.
- PARCELNO – identifying each property.
- SALE_PRC – in “\$”
- LND_SQFOOT – the land the property is accompanied by in square feet.

- TOT_LVG_AREA – the living area of the property in square feet.
- SPEC_FEAT_VAL – value of added property enhancements in “\$”.
- RAIL_DIST – the distance from the closest train service in feet.
- OCEAN_DIST - the distance from the closest ocean/sea in feet.
- WATER_DIST - the distance from the closest water pool in feet.
- CNTR_DIST – the distance to the closest business area in feet.
- SUBCNTR_DIST - the distance to the closest subcentre in feet.
- age – age of the house in the asset.
- avno60plus – measuring the noise generated by air transportation above the appropriate levels.
- month_sold – month when the property was sold in 2023.
- structure_quality – quality of building of the house.

It is also visible that the variables are stored as numerical (namely “int” – representing integers and “num” – representing numerical doubles). However, it is a good rule of thumb for some of the variables to be converted to categorical (month_sold, structure_quality, PARCELNO) as they represent nominal or ordinal codes, which can evade confusing data splits, distance estimations and better the transparency of the model’s decision-making (Kuhn *et al.*, 2019; Wickham *et al.*, 2023).

```
'data.frame': 13932 obs. of 17 variables:
 $ LATITUDE      : num  25.9 25.9 25.9 25.9 25.9 ...
 $ LONGITUDE     : num  -80.2 -80.2 -80.2 -80.2 -80.2 ...
 $ PARCELNO      : num  6.22e+11 6.22e+11 6.22e+11 6.22e+11 6.22e+11 ...
 $ SALE_PRC      : num  440000 349000 800000 988000 755000 630000 1020000 850000 250000 1220000 ...
 $ LND_SQFOOT    : int   9375 9375 9375 12450 12800 9900 10387 10272 9375 13803 ...
 $ TOT_LVG_AREA  : int   1753 1715 2276 2058 1684 1531 1753 1663 1493 3077 ...
 $ SPEC_FEAT_VAL : int    0 0 49206 10033 16681 2978 23116 34933 11668 34580 ...
 $ RAIL_DIST     : num   2816 4359 4413 4585 4063 ...
 $ OCEAN_DIST    : num  12811 10648 10574 10156 10837 ...
 $ WATER_DIST    : num   348 338 297 0 327 ...
 $ CNTR_DIST     : num  42815 43505 43530 43798 43600 ...
 $ SUBCNTR_DI    : num  37742 37341 37329 37423 37551 ...
 $ HWY_DIST      : num  15955 18125 18201 18514 17903 ...
 $ age           : int    67 63 61 63 42 41 63 21 56 63 ...
 $ avno60plus    : int    0 0 0 0 0 0 0 0 0 0 ...
 $ month_sold    : int    8 9 2 9 7 2 2 9 3 11 ...
 $ structure_quality: int    4 4 4 4 4 4 5 4 4 5 ...
```

Figure 4 - Dataset Structure.

From **Figures 5 & 6** can be observed information regarding the mean, median, max, standard deviation values. The analysis and insights of which have been summarised in **Table 3**.

LATITUDE	LONGITUDE	PARCELNO	SALE_PRC	LND_SQFOOT	TOT_LVG_AREA	SPEC_FEAT_VAL
Min. :25.43	Min. :-80.54	Min. :1.020e+11	Min. : 72000	Min. : 1248	Min. : 854	Min. : 0
1st Qu.:25.62	1st Qu.: -80.40	1st Qu.:1.079e+12	1st Qu.: 235000	1st Qu.: 5400	1st Qu.:1470	1st Qu.: 810
Median :25.73	Median :-80.34	Median :3.040e+12	Median : 310000	Median : 7500	Median :1878	Median : 2766
Mean :25.73	Mean :-80.33	Mean :2.356e+12	Mean : 399942	Mean : 8621	Mean :2058	Mean : 9562
3rd Qu.:25.85	3rd Qu.: -80.26	3rd Qu.:3.060e+12	3rd Qu.: 428000	3rd Qu.: 9126	3rd Qu.:2471	3rd Qu.: 12352
Max. :25.97	Max. :-80.12	Max. :3.660e+12	Max. :2650000	Max. :57064	Max. :6287	Max. :175020
RAIL_DIST	OCEAN_DIST	WATER_DIST	CNTR_DIST	SUBCNTR_DI	HWY_DIST	age
Min. : 10.5	Min. : 236.1	Min. : 0	Min. : 3826	Min. : 1463	Min. : 90.2	Min. : 0.00
1st Qu.: 3299.4	1st Qu.:18079.3	1st Qu.: 2676	1st Qu.: 42823	1st Qu.: 23996	1st Qu.: 2998.1	1st Qu.:14.00
Median : 7106.3	Median :28541.8	Median : 6923	Median : 65852	Median : 41110	Median : 6159.8	Median :26.00
Mean : 8348.5	Mean :31691.0	Mean :11960	Mean : 68490	Mean : 41115	Mean : 7723.8	Mean :30.67
3rd Qu.:12102.6	3rd Qu.:44310.7	3rd Qu.:19200	3rd Qu.: 89358	3rd Qu.: 53949	3rd Qu.:10854.2	3rd Qu.:46.00
Max. :29621.5	Max. :75744.9	Max. :50400	Max. :159977	Max. :110554	Max. :48167.3	Max. :96.00
avno60plus	month_sold	structure_quality				
Min. :0.00000	Min. : 1.000	Min. :1.000				
1st Qu.:0.00000	1st Qu.: 4.000	1st Qu.:2.000				
Median :0.00000	Median : 7.000	Median :4.000				
Mean :0.01493	Mean : 6.656	Mean :3.514				
3rd Qu.:0.00000	3rd Qu.: 9.000	3rd Qu.:4.000				
Max. :1.00000	Max. :12.000	Max. :5.000				

Figure 5 - Dataset Summary.

Data Frame Summary

housing_data

Dimensions: 13932 x 17

Duplicates: 0

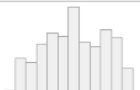
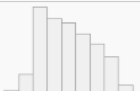



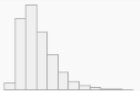

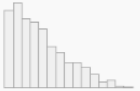
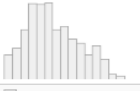
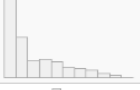
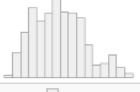
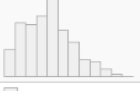
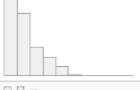
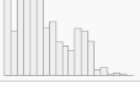
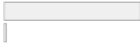
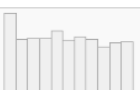

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	LATITUDE [numeric]	Mean (sd) : 25.7 (0.1) min ≤ med ≤ max: 25.4 ≤ 25.7 ≤ 26 IQR (CV) : 0.2 (0)	13776 distinct values		13932 (100.0%)	0 (0.0%)
2	LONGITUDE [numeric]	Mean (sd) : -80.3 (0.1) min ≤ med ≤ max: -80.5 ≤ -80.3 ≤ -80.1 IQR (CV) : 0.1 (0)	13776 distinct values		13932 (100.0%)	0 (0.0%)
3	PARCELNO [numeric]	Mean (sd) : 2.356496e+12 (1.19929e+12) min ≤ med ≤ max: 102000801020 ≤ 3.0403e+12 ≤ 3.66017e+12 IQR (CV) : 1.98101e+12 (0.5)	13776 distinct values		13932 (100.0%)	0 (0.0%)
4	SALE_PRC [numeric]	Mean (sd) : 399941.9 (317214.7) min ≤ med ≤ max: 72000 ≤ 310000 ≤ 2650000 IQR (CV) : 193000 (0.8)	2111 distinct values		13932 (100.0%)	0 (0.0%)
5	LND_SQFOOT [integer]	Mean (sd) : 8620.9 (6070.1) min ≤ med ≤ max: 1248 ≤ 7500 ≤ 57064 IQR (CV) : 3726.2 (0.7)	4696 distinct values		13932 (100.0%)	0 (0.0%)
6	TOT_LVG_AREA [integer]	Mean (sd) : 2058 (813.5) min ≤ med ≤ max: 854 ≤ 1877.5 ≤ 6287 IQR (CV) : 1001 (0.4)	2978 distinct values		13932 (100.0%)	0 (0.0%)
7	SPEC_FEAT_VAL [integer]	Mean (sd) : 9562.5 (13891) min ≤ med ≤ max: 0 ≤ 2765.5 ≤ 175020 IQR (CV) : 11542.2 (1.5)	7583 distinct values		13932 (100.0%)	0 (0.0%)
8	RAIL_DIST [numeric]	Mean (sd) : 8348.5 (6178) min ≤ med ≤ max: 10.5 ≤ 7106.3 ≤ 29621.5 IQR (CV) : 8803.2 (0.7)	13235 distinct values		13932 (100.0%)	0 (0.0%)
9	OCEAN_DIST [numeric]	Mean (sd) : 31691 (17595.1) min ≤ med ≤ max: 236.1 ≤ 28541.8 ≤ 75744.9 IQR (CV) : 26221.2 (0.6)	13617 distinct values		13932 (100.0%)	0 (0.0%)
10	WATER_DIST [numeric]	Mean (sd) : 11960.3 (11933) min ≤ med ≤ max: 0 ≤ 6922.6 ≤ 50399.8 IQR (CV) : 16524.2 (1)	13218 distinct values		13932 (100.0%)	0 (0.0%)
11	CNTR_DIST [numeric]	Mean (sd) : 68490.3 (32008.5) min ≤ med ≤ max: 3825.6 ≤ 65852.4 ≤ 159976.5 IQR (CV) : 46535.2 (0.5)	13682 distinct values		13932 (100.0%)	0 (0.0%)
12	SUBCNTR_DI [numeric]	Mean (sd) : 41115 (22161.8) min ≤ med ≤ max: 1462.8 ≤ 41109.9 ≤ 110553.8 IQR (CV) : 29953.1 (0.5)	13642 distinct values		13932 (100.0%)	0 (0.0%)
13	HWY_DIST [numeric]	Mean (sd) : 7723.8 (6068.9) min ≤ med ≤ max: 90.2 ≤ 6159.8 ≤ 48167.3 IQR (CV) : 7856.1 (0.8)	13213 distinct values		13932 (100.0%)	0 (0.0%)
14	age [integer]	Mean (sd) : 30.7 (21.2) min ≤ med ≤ max: 0 ≤ 26 ≤ 96 IQR (CV) : 32 (0.7)	96 distinct values		13932 (100.0%)	0 (0.0%)
15	avno60plus [integer]	Min : 0 Mean : 0 Max : 1	0 : 13724 (98.5%) 1 : 208 (1.5%)		13932 (100.0%)	0 (0.0%)
16	month_sold [integer]	Mean (sd) : 6.7 (3.3) min ≤ med ≤ max: 1 ≤ 7 ≤ 12 IQR (CV) : 5 (0.5)	12 distinct values		13932 (100.0%)	0 (0.0%)
17	structure_quality [integer]	Mean (sd) : 3.5 (1.1) min ≤ med ≤ max: 1 ≤ 4 ≤ 5 IQR (CV) : 2 (0.3)	1 : 179 (1.3%) 2 : 4110 (29.5%) 3 : 16 (0.1%) 4 : 7625 (54.7%) 5 : 2002 (14.4%)		13932 (100.0%)	0 (0.0%)

Figure 6 - Detailed Dataset Summary.

Table 3 - Data Summary Analysis and Insights Table (Kuhn et al., 2019; Wickham et al., 2023).

Variable (examples)	Mean vs Median	Range (Min / Max)	What it suggests
SALE_PRC	Mean - \$399 k Median - \$310 k	\$72 k to \$2.65 M	Mean comfortably above the median suggesting a right-skewed distribution typical of housing prices. Expect many properties clustered below \$500 k and a tail of high-value sales; potential need for log-transform in regression.
LND_SQFOOT	Mean - 8 621 Median - 7 500	1 248 to 57 064	Long right tail (large lots). The wide range implies heteroscedasticity risk; consider scaling or winsorising extremes.
TOT_LVG_AREA	Mean - 2 058 Median - 1 878	854 to 6 287	Similar skew to land size but with a smaller absolute spread.
SPEC_FEAT_VAL	Mean - 9 562 Median - 2 766	0 to 175 200	High skew — many houses have little or no special-feature value while a few have pools/outbuildings worth >\$150 k. Zero-inflation may need special treatment (e.g., two-part modelling or transformation).
Distance metrics (RAIL_DIST, OCEAN_DIST, CNTR_DIST, etc.)	Means always much larger than medians	E.g., HWY_DIST 97 to 48 167	Right-skew indicates most properties are fairly close to the amenity/noise source, with a minority located far away. These variables span orders of magnitude—standardisation will help many algorithms.
age	Mean - 31 yrs Median - 26 yrs	0 to 96	Fairly symmetric (mean close to 3rd quartile midpoint) but shows a handful of very old properties nearing 100 years.
avno60plus	Median = 0, 3Q = 0	0 to 1	Imbalanced and should be treated as categorical.
month_sold	Median = 4 (April)	1 to 12	present.
structure_quality	Mean - 5.7 Median = 5	1 to 10	Near-symmetry around the mid-scale value, validating its use as an ordered predictor.

3.3. Data Exploration (Preliminary Analysis).

From **Figure 7** it is clearly visible that the spread of property costs is skewed significantly to the right. Thus, the majority of abode deals cluster around the range of 200 to 450 thousand dollars. The data tail is stretching continuously beyond two million dollars.

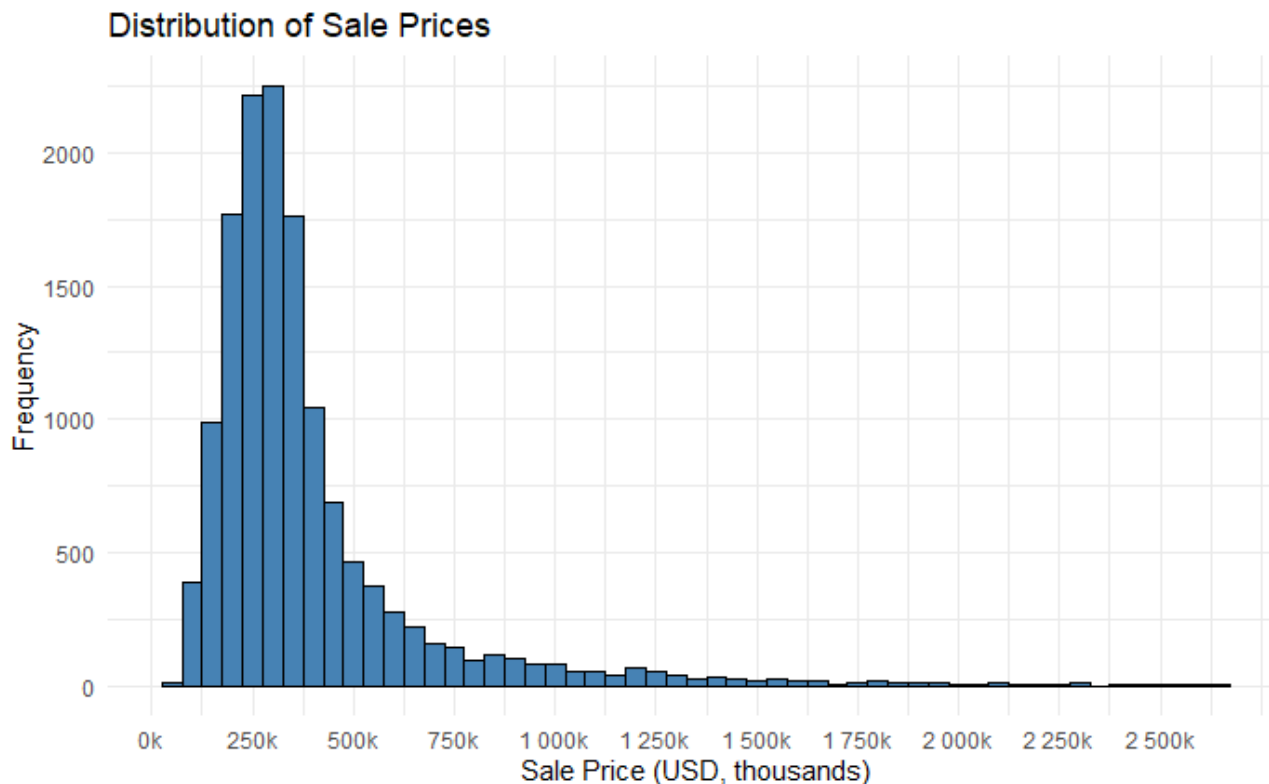


Figure 7 - Histogram for Distribution of Sale Prices.

In **Figure 8** is portrayed the outlier occurrence, which is showing a lot of instances that surface above the upper whisker that is approximately reflecting the nine hundred-thousand-dollar mark. The midspread of data approximates between the range of 235 – 430k \$.

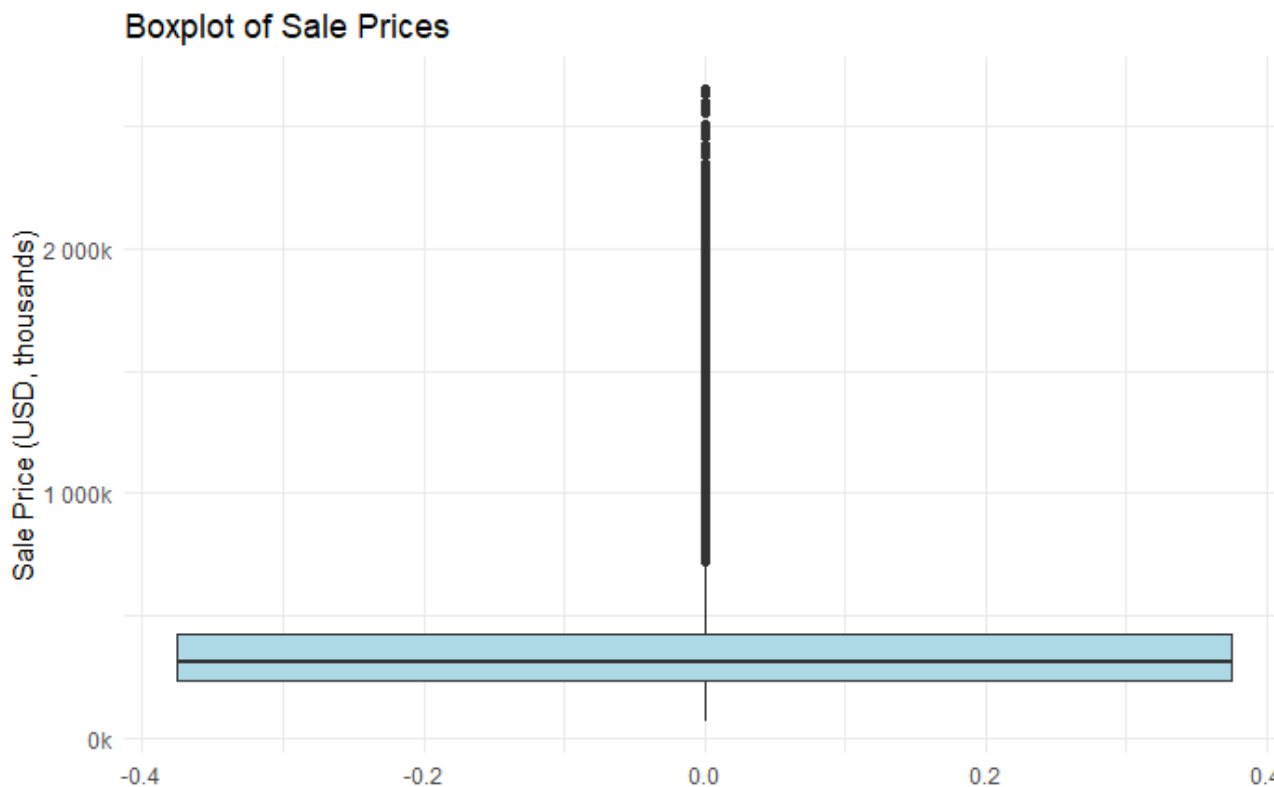


Figure 8 - Box Plot of House Sale Prices.

Figure 9 depicts a correlational heatmap in order to identify the relationships among different data features. The **strong positive associations** can simultaneously be observed between the LND_SQFOOT and SALE_PRC ($r=0.67$), LND_SQFOOT AND SPEC_FEAT_VAL ($r=0.51$), and CNTR_DIST to SUBCNTR_DI ($r=0.77$). Hence, the land area applies the biggest influence on the property cost. Within parcels that have larger land plots seem to be a higher number of additional feature investments. On the other hand, the two aforementioned metrics for plot proximity to centres seem to be almost equal. **Modest correlation** seems to be present between the total living area of a property and the sale price ($r=0.23$), and between the age of the asset and its selling cost ($r=-14$). It looks like the living

area size is moderately affecting the abode cost and newer parcels exhibit a slightly higher trade return.

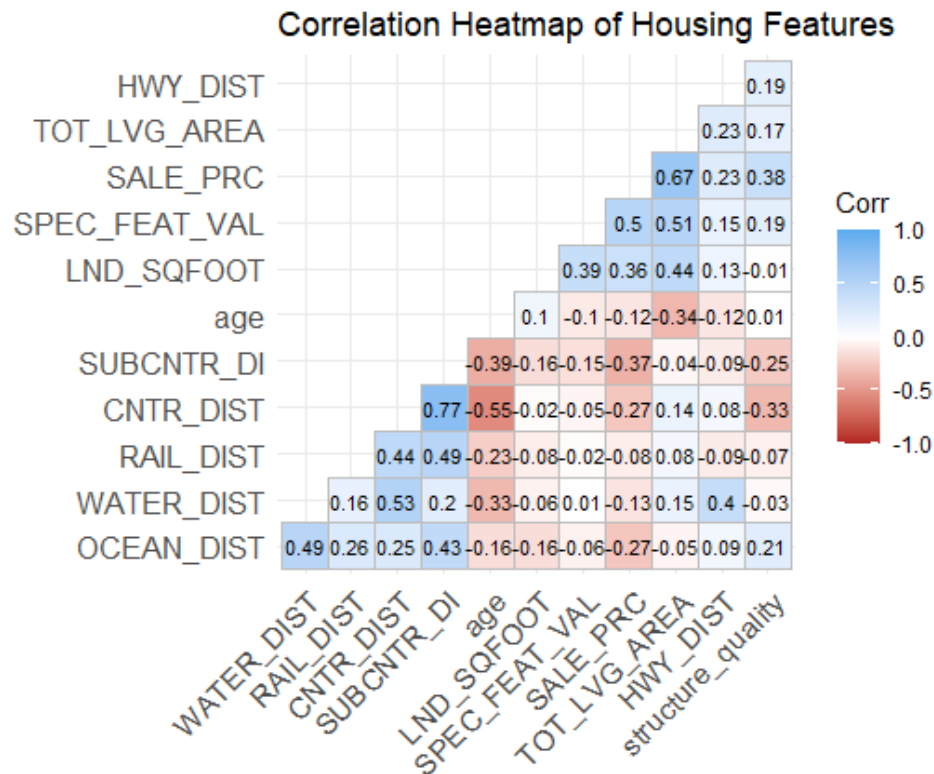


Figure 9 - Correlational Heatmap of Housing Features.

3.4. Data Quality Verification.

From the output presented in **Figure 10** it is clearly visible that every column is completed, thus there are no missing values present in the data frame. Hence, the need of imputation is evaded at this phase.

LND_SQFOOT	LATITUDE	LONGITUDE	PARCELNO	SALE_PRC
	TOT_LVG_AREA			
0	0	0	0	0
	SPEC_FEAT_VAL	RAIL_DIST	OCEAN_DIST	WATER_DIST
	CNTR_DIST	SUBCNTR_DI		
0	0	0	0	0
	HWY_DIST	age	avno60plus	month_sold
	structure_quality			
0	0	0	0	0

Figure 10 - Missing Values per Column Check.

During the check for out-of-range data points in the “age” and “structure_quality” columns (**Figure 11**), it is observed that all present data inputs can be classified as realistic and valid. Meaning, that newly built assets are represented by age value of “0” and the oldest recorded home has been built 96 years ago. On the other hand, the structure quality seems to be ordinally scaled between the values of “1” and “5”.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	14.00	26.00	30.67	46.00	96.00
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.000	4.000	3.514	4.000	5.000

Figure 11 - Checking for out-of-range values in the “age” and “structure_quality” columns.

When checking for duplicate occurrences in the “PARCELNO” column (the primary key domain), the investigation shows that 152 out of 13,932 records have the same identifier surfacing between 2 and 3 times, thus hinting about multiple trade deals regarding the same property thought the year of 2023, rather than data input mistakes. This approximately estimates to 1.1 % of all data records (**Figure 12**).

PARCELNO <dbl>	n <int>
1.311305e+11	2
1.311307e+11	2
1.311404e+11	2
1.312301e+11	2
1.312402e+11	2
1.312402e+11	2
1.312402e+11	3
1.312502e+11	2
1.312704e+11	2
1.313301e+11	2

1-10 of 152 rows

Previous **1** 2 3 4 5 6 ... 16 Next

Figure 12 - Multiple Occurances of the Same Parcel Number in the Dataset.

While a check for internal consistency is performed there are no conflicting values that can be identified across the distance or monetary fields, and no negative areas or prices (**Figure 13**).

0 rows | 1-9 of 17 columns

Figure 13 - Checking for Potential Data Anomalies.

4. Data Preparation Phase (Data Munging).

In this step of the project, the emphasis falls on getting the data frames ready for the following Modelling step. Thus, bringing them to a good working standard, so satisfactory outcomes can be achieved. It is a well-known fact that such projects are heavily dependent on the data quality that is fed into the chosen models, more accurately 80% of the project's achievements are said to be dependent on the latter (Putler and Krider, 2012).

4.1. Selecting the data.

In this step it has been decided to remove the “**LATITUDE**”, “**LONGITUDE**”, and “**PARCELNO**” columns. The first two are considered irrelevant and not needed to the scope of the project as it has been shared by the stakeholders that the data is coming from the Bolton region, and the project is supposed to deliver price forecasts for the same one. Hence, knowing the precise location of the property would not add any predictive value to the models. It would have been another case if the data held records regarding the neighbourhood in which the asset is situated, as this is a well-known factor that affects the price. Then the columns might have been retained. “**PARCELNO**” is also considered irrelevant because it acts as an identifier to the property with no direct effect to the target value of the sale price. Also, since it is directly involved in identifying a specific abode, it might create GDPR issues by allowing for identification of the property or its owners. Thus, its removal. In such a manner dimensionality and data spillage issues can be tackled and evaded (Zhang *et al.*, 2021; Loukides *et al.*, 2018; Kuhn *et al.*, 2019).

4.2. Data Cleaning.

As discussed in the previous phase of the project, all domains have been confirmed to be complete and lack any missing fields. Therefore, the need for imputation is not a requisite. Additionally, while the dataset has been assessed, the emergence of the possibility of whether to retain all duplicating sales for the same asset in the year or retain only the latest

sale records appeared (152 occurrences). Therefore, a decision has been made to retain all records and to allow the models to observe how the price evolves during the year, rather than aiming for a single evaluation (Kuhn *et al.*, 2019).

4.3. Construct the Data (Feature Engineering).

During this stage the “SALE_PRICE” column was log transformed and the “month_sold”, “avno60plus”, and “structure_quality” were converted to factors, so the models can treat them appropriately. It is important to note that both the raw sale price was kept and a log transformed one was added as a column. This was done so the variety of models that would be trained during the Modelling phase can be aimed at the target data they would work better on. During the log transformation was applied a log of 10. This has been administered because the technique betters homoscedasticity and brings in normality to the heavily right skewed data. Thus, improving the application of regression models for the task. Additionally, the conversion to factors has been implemented, so algorithms would not take them for empirical distances (James *et al.*, 2021).

4.4. Integration of Data (Formatting for Modelling).

In this step the data frame has been scaled. In such a manner, the data is warranted to be cantered which will guarantee that where architectures that are sensitive to deviations in magnitude are going to be affected equally by the continuous predictors (Han *et al.*, 2011).

5. Modelling Phase.

This stage of the project is centring around the selection, building, testing and evaluating the models' performance, so the best fit for the project purposes can be selected (Putler and Krider, 2012).

5.1. Selection of Model Techniques.

5.1.1. Multiple Linear Regression (MLR).

MLR is an ML model that is well-founded and tested in predictive analytics. It is one of the benchmark models. It presents fully translucent coefficients and well-researched statistical properties. By assuming additive and approximately linear effects, it allows stakeholders to quantify how each square-foot of land or each additional year of age alters the sale price, controlling for other variables. The abode market can have occurrences of non-linearity, keeping an MLR baseline can serve as a starting comparison point that more sophisticated models can be tested against (James *et al.*, 2021).

5.1.2. Support-Vector Regression (SVR).

It uplifts the support-vector idea to targets that are continuous, where it also betters the margin-based loss that yields strong generalisation performance even in high-dimensional spaces. There are three known types of the model (Scholkopf and Smola, 2001):

- A **quick linear conceptualisation** (Liblinear back end) – aimed at large-scale baseline accuracy.
- **Polynomial** – can model subtle curvature such as reducing land-size benefits.
- **Radial-basis-function** – able to assess complex and localised patterns in data.

In such a manner can be observed and investigated how the growing kernel complexity may better foretelling accuracy while warranting reduction in over-fitting through the cross-validation process (Scholkopf and Smola, 2001).

5.1.3. Decision Tree.

Such approaches separate the predictor space via a number of axis-aligned splits, delivering a sophisticated and easily interpretable collection of “if-then” rules. Thus, for abode markets, the model produces a threshold effect. Although, singular trees can be unstable, their observational cleanliness is of utmost importance when outcomes need to be delivered to non-technical stakeholders and for revealing candidate interaction terms, which can outline following feature engineering (Fan *et al.*, 2006).

5.1.4. Random Forest.

They lump hundreds of uncorrelated decision trees, each one of which is grown on a bootstrap sample and a random subset of predictors. In such a manner bias is kept at low levels and variance is diminished. This allows them to intercept intricate non-linear interactions like the common effect the land area, structure quality, and multiple distance measures would have. This allows for a reduced amount of hyper-parameter tuning. They are also historically known to outperform other models, which makes them a good fit for the project’s purposes (Breiman, 2001).

5.2. Testing Design.

Since the datasets consists of almost fourteen thousand records a good split ratio of 80% to 20% has been applied. Where the larger percentage goes for the set that is utilised for training of the models and the lesser for the testing (Kuhn *et al.*, 2019). As for the error metrics that will be used to compare the models performance, the following were selected (Karunasingha, 2022; James *et al.*, 2021):

- RMSE – would castigate sufficient errors in dollars. Would prove useful, because the dataset has outliers worth millions.
- MAE – provides a scale-consistent “typical” error, less sensitive to extreme values.
- R squared – communicates variance explained, easily understood by business stakeholders.

5.3. Model Building.

All of the chosen models were constructed, and all three variants of SVR have been utilised. The **MLR** is using the log-transformed target column to mitigate the right-skew of the data. The three kernels of the **SVR** model have also been utilised, where the Liblinear backend with a compact cost grid of 0.25 – 16 has been used for computational efficiency. The polynomial variation (with degrees 2 – 3) and RBF kernels have been tuned over cost and kernel parameters to capture moderate and highly- non-linear relationships respectively. The **Decision Tree** where a rpart tree was flowered on the raw price scale with a value of 20 complexity parameter (cp) grid. Five-fold cross-validation and a verbose setting provided live feedback and warrantied the selected cp balanced bias and variance. With **RF** five hundred trees were built with mtry selected by caret’s default optimisation. A progress ticker of fifty confirmed build completion within three minutes on six logical CPU cores. RF was trained on the untransformed price to exploit its scale-invariant splitting criterion (James *et al.*, 2021; Fan *et al.*, 2006; Scholkopf and Smola, 2001; Breiman, 2001).

The complete set of empirical predictors except the two target columns were centred and scaled. Hyper-parameter tuning for every model used 5-fold cross validation repeated across the specified grids, and the best setting by lowest RMSE was refit on the full training partition. Each model was saved in a separate R object.

5.4. Assessing of Model Performance.

When assessing the performance of the models it should be considered that lower RMSE and MAE indicate smaller dollar-scaled prediction error. Additionally, where R squared is higher would rectify that more variance in sale price is expected (Witten *et al.*, 2017).

From **Table 4 and Figure 14** can be observed that the Linear regression model that has been utilised as a baseline has the largest errors, closely followed by the SVR-Linear model. The SVR-Poly one performed reasonably well by successfully capturing some of the non-linearity. It is closely comparable in performance with the SVR-RBF variation. The Decision Tree on the other hand, exhibited high error and low explanatory power. The **RF model** proves to be the best performing architecture of all tested. It outmanoeuvres the rest by portraying the lowest RMSE (approximately \$99k) and lowest MAE (around \$47k), with the highest r squared value (0.89).

The ensemble of decorrelated decision trees effectively outlines intricate, non-linear relationships in conjunction with a controlled over-fitting. Thus, it provides with the most accurate and invariant foretelling on the unseen test data.

Table 4 - Test Set Evaluation Result Table.

Model <chr>	RMSE <dbl>	MAE <dbl>	R2 <dbl>
Linear (log)	167138.60	84191.61	0.7168339
SVR-Linear (log)	168803.81	83778.20	0.7033918
SVR-Poly (log)	111119.42	53140.57	0.8648832
SVR-RBF (log)	116889.59	53442.84	0.8517452
Decision Tree	166004.11	97845.07	0.7058634
Random Forest	98936.59	47241.94	0.8922191

6 rows

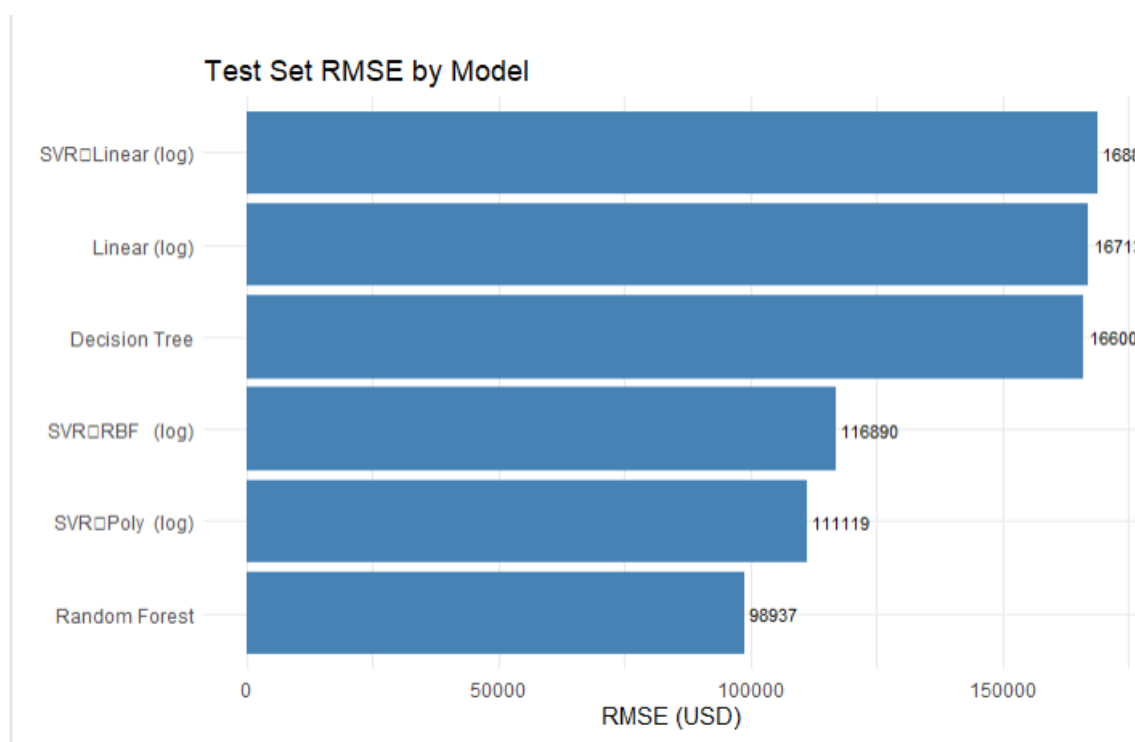


Figure 14 - RMSE Comparison Result Bar Chart.

6. Evaluation Phase.

This step of the process is focused on a wider range than the previous one, more precisely to understand whether the delivered solution fulfils the business goals and criteria (Putler and Krider, 2012).

6.1. Evaluation of Results Against Business Goals.

The accuracy target that the stakeholders want seems to be met by the RF model that outperforms all of the other tried solutions that can clearly be seen from Figure 14. The 5-fold cross-validation confirms stable and **robust** performance of the best solution where small hyper-parameter deviations did not have a negative effect on its RMSE. Although, RF is less **transparent** than a single tree, variable-importance plots show land size, special-feature value and distance to CBD as the three dominant predictors. Hence, they act as interpretability enhancers and can be easily communicated even to non-technical stakeholders. RF also introduces almost minimal latency, thus it is **very cost effective** and can deploy into production at a significantly lower cost, if compared to deep learning solutions. Respectively, it can be concluded that RF has been approved as the optimal model.

6.2. Review of Process.

Table 5 outlines a brief review of process that has been performed, which showed that no pivotal project iterations were overlooked. However, future iterations could delve into gradient-boosting ensembles (e.g., XGBoost) as possible improvements.

Table 5 - Review of Process Table.

CRISP-DM phase	Key checks	Outcome
Business & Data Understanding	Objectives explicitly tied to valuation use-case; data quality verified	✓
Data Preparation	All numeric predictors scaled; categorical recoded; duplicates handled; targets retained on raw & log scale	✓
Modelling	Five algorithms tuned with cross-validation; progress logged; results stored	✓
Evaluation	RF clearly best; interpretation and business alignment documented	✓

6.3. Prediction Generation for a Specified Property.

As per stakeholder's request a prediction has been delivered with specified criteria:

PARCELNO: 728980145245, LND_SQFOOT: 11247, TOTLVGAREA: 4552, SPECFEATVAL: 2105, RAIL_DIST: 4871.9, OCEAN_DIST: 18507.2, WATER_DIST: 375.8, CNTR_DIST: 43897.9, SUBCNTR_DI: 40115.7, HWY_DIST: 41917.1, age: 42, avno60plus: 0, structure_quality: 5, month_sold: 8.

This has outputted as **\$1 453 318**. Although, the median in price is around \$310k the prediction made still fits below the maximum level of \$2.65 million. Also, the assets attributes suggest that a larger living area is present (4552 sq.ft.) and a big lot size (11247 sq. ft.) with a small distance to water bodies and a modest distance to rail and highway services. The structure quality in the prompt also supposes this is a high-end property. Thus, the valuation satisfies the business criterion and can be presented to the stakeholders.

6.4. Next Steps.

6.4.1. Deployment Plan

The trained model can be saved alongside the preprocessing steps, and they can be deployed, so analysts can utilise them via an API and receive direct valuations (James *et al.*, 2021).

6.4.2. Monitoring and Maintenance.

The RF model can be re-trained quarterly with new data that becomes available to better reflect errors and changes in the market.

6.4.3. Future Improvements.

As mentioned previously gradient-boosting and the addition of hedonic spatial-lag features can be added to capture neighbourhood spill-over effects (James *et al.*, 2021).

7. Conclusion

The current project applied the CRISP-DM project life cycle to foretell abode costs in the Bolton area. Starting with understandable business goals-to achieve a model accurate enough to support investment decisions, the project progressed through data understanding, thoughtful preparation, training and testing multiple models, and a deep evaluation.

The exploratory analysis revealed heavy price skew, multicollinearity among distance metrics, and a handful of duplicate parcel sales. These pushed forward some aimed cleaning (log transformation, selective scaling, factor recoding) and the extraction of irrelevant geographical coordinates. Six algorithms have been tested via a five-fold cross-validation on an 80/20 data split. The Random Forest solution proved to be the best performing.

The model of choice was then tested with a pre-set requirements by the stakeholders and proved to be an accurate prediction maker by valuating a high-end property to the estimate of \$ 1.45m. Further, the model is ready for deployment and can be implemented on an API to provide analysts with means of immediate valuation. The project concluded with some propositions for improvements like looking into gradient-based boosting.

8. Bibliography

Breiman, L. (2001) *Random Forests*. 45, 5–32.

Fan, G.Z. et al. (2006) Determinants of House Price: A Decision Tree Approach. *Urban Studies*, 43(12), 2301–2316. Sage PublicationsSage UK: London, England. [Accessed: 18 April 2025].

Fawcett, T. & Provost, F. (2013) Data Science for Business. In: Loukides, M. & Blanchette, M. (eds.) 2013, 1st ed. [Online]. O'Reilly Media, Inc. Available at: https://www.researchgate.net/publication/256438799_Data_Science_for_Business [Accessed: 17 April 2025].

Hamizah Zulkifley, N. et al. (2020) House Price Prediction using a Machine Learning Model: A Survey of Literature. *Modern Education and Computer Science*, 6, 46–54. [Accessed: 16 April 2025].

Han, J. et al. (2011) *Data Mining: Concepts and Techniques (Google eBook)*. [Online]. Elsevier. Available at: <http://books.google.com/books?id=pQws07tdpjoC&pgis=1> [Accessed: 17 April 2025].

James, G. et al. (2021) *An Introduction to Statistical Learning*. [Online]. New York, NY: Springer US. Available at: doi:10.1007/978-1-0716-1418-1 [Accessed: 16 April 2025].

Karunasingha, D.S.K. (2022) Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences*, 585, 609–629. Elsevier. [Accessed: 18 April 2025].

Kuhn, M. et al. (2019) Feature engineering and selection: A practical approach for predictive models. In: Informa UK Limited *The American Statistician*. [Online]. Informa UK Limited. Available at: doi:10.1080/00031305.2020.1790217 [Accessed: 17 April 2025].

Larose, C.D. & Larose, D.T. (2019) *Data science using Python and R*. [Online]. John Wiley & Sons, Inc. Available at: <https://learning.oreilly.com/library/view/data-science-using/9781119526810/> [Accessed: 16 April 2025].

Loukides, Mike. et al. (2018) *Ethics and Data Science*. [Online]. O'Reilly Media, Inc. Available at:

https://bolton.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwpV1NS8NAEB1q68GTiopflRzEW5rNbrKbQIXQtggeilYKXiy7yUTENoV-ePS3O5uPQr148JoITvZTN5s5r0HIHiLub9yAhXNKIXRiUwJlvuJL9CEiqdGCI5oLN95eCden8LRsxjWoF9RY0rrhPeW9amcTNa_UXOxB7t37n17uWpT23l8o3bk3er5FU71x-T6bbUFDYL3ihZ7YzDo9h6rhWbFWOINlznRqzxTVPP1ZWsdqhOqWAtGiPz701_FxbrW7PN1ut2EzOzJtAl_2tT0PH_89iDBlrq wz7UMDuAy85Uzz9vioZ4R2eJ09VL7ZTJoOPlo4cw6vde7h_c0IPBNT4Vo8w1mBJoiZWKY oU-E3Eg6UCS-

FwZAKc6ZDpEX1GlpGbSqMAoAmUiMDrW3FYbR1DPZhkegxNQbjKBYJQiU-t6FUlpYmQy
[Accessed: 18 April 2025].

Putler, D.S. & Krider, R.E. (2012) A Process Model for Data Mining—CRISP-DM. In: Chapman and Hall/CRC *Customer and Business Analytics*. [Online]. Chapman and Hall/CRC. Available at: doi:10.1201%2Fb12040-8 [Accessed: 16 April 2025].

Scholkopf, Bernhard. & Smola, A.J.. (2001) Learning with Kernels. In: MIT Press *Adaptive Computation and Machine Learning series, 2001, 1* Adaptive Computation and Machine Learning series. [Online]. MIT Press. Available at: <https://ebookcentral.proquest.com/lib/bolton/detail.action?docID=3338886&pq-origsite=summon> [Accessed: 18 April 2025].

Shmueli, G. et al. (2017) *Data mining for business analytics: concepts, techniques, and applications in R*. [Online]. Available at: <https://books.google.com/books?hl=en&lr=&id=ETwuDwAAQBAJ&oi=fnd&pg=PR19&dq=at+a+Mining+for+Business+Analytics:+Concepts,+Techniques,+and+Applications+in+R&ots=2PXcrq5WML&sig=Ce4FLe8vycr3Wex2WuUHv4Qv1XY> [Accessed: 17 April 2025].

Wickham, H. et al. (2023) R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. In: O'Reilly 2023, *2nd edition*. [Online]. O'Reilly. Available at: <https://www-oreilly-com.ezproxy.bolton.ac.uk/library-access/?next=/library/view/~/9781492097396/&email=%2Bhp9P49CamjLDmv5eqqqg8NEqj5sZju0&tstamp=1744914562&id=83A64946D3B073B9182E592F6F852F5296216C9D> [Accessed: 17 April 2025].

Witten, I.H., et al. (2017) Data Mining: Practical Machine Learning Tools and Techniques,. In: Elsevier Science & Technology *Data Mining*. 4th ed. [Online]. Elsevier Science & Technology. Available at: doi:10.1016/C2009-0-19715-5 [Accessed: 16 April 2025].

Zhang, D. et al. (2021) Regional housing price dependency in the UK: A dynamic network approach. *Urban Studies*, 58(5), 1014–1031. SAGE Publications Ltd. [Accessed: 18 April 2025].

9. Word Count

3649 words

10. GAI Declaration

I declare that no GAI was used for the development of this work.

The software used is:

- Microsoft Word and its grammar/spelling corrector
- Microsoft Excel – for the development of comparison tables and graphs
- Mendeley Reference Manager
- Google, Google Scholar and Discover@Bolton to narrow down and uplift research

11. Appendices

11.1. List of Abbreviations Used

1. **AI** - Artificial Intelligence
2. **CNN** – Convolutional Neural Network
3. **CRISP-DM** - Cross-Industry Process for Data Mining
4. **DA** – Data Analysis
5. **DL** – Deep Learning
6. **DM** – Data Mining
7. **DS** – Data Science
8. **EDA** – Exploratory Data Analysis
9. **ML** – Machine Learning
10. **MLR** – Multiple Linear Regression
11. **MSE** – Mean Squared Error
12. **NN** – Neural Network
13. **RMSE** – Root Mean Squared Error
14. **SVR** – Support-Vector Regression