Rotterdam School of Management
Erasmus University

Group Project

Real Estate Price Prediction

A consultancy project for RealtyCheck

Team 7

Mila Liu	533765	Data preparation and modeling steps
Lubna Begum	505263	Business understanding and data understanding steps
Huaying Zhu	633752	Weight of variables importance, result discussion
Elizabeth Feng	589855	Model evaluation and deployment steps
Khoi Pham	523755	Building prediction model

Course: Big Data Management & Analytics (BM04BIM)

MSc Business Information Management

Rotterdam School of Management, Erasmus University

Table of Contents

Part I – System Proposal

1. Introduction

The first part of this report provides a structural overview of the automated system that helps RealtyCheck identify attractive properties and advise on their value. We sequentially address the company's concerns through six steps of the CRISP-DM cycle. The first two steps, Business understanding and Data understanding, identify relevant business objectives and explain how the automated system helps achieve them. In the third step, we identified the important metrics to be considered when assessing a property and advised on the type and sources to collect such information. Afterward, we explain why the Regression tree is our selected model. In this Modelling step, we also advise on the collection frequency of different data types. Next, we clarify the system performance assessment in the Evaluation step. And finally, in the Deployment step, we give suggestions on how the company may adopt the Scrum process to monitor and update the system in practice.

2. CRISP-DM cycle

2.1 Business understanding

This section presents the potential added-value and strategic advantages of the automated system to RealtyCheck through the first stage in the CRISP-DM cycle, business understanding. RealtyCheck business objective is to make smarter real-estate investments by taking advantage of market fluctuations, with a focus on the Rotterdam area. Real estate valuation has always been a complex and time-consuming process, as house prices are volatile and usually influenced by multiple social and economic metrics (Braun et al., 2022). A system that helps identify attractive real estate properties before market listing promisingly benefits firms like RealtyCheck in the following ways. First, RealtyCheck is able to identify the precise valuations of properties and make the correct decision. After the market listing of the property, the system shall help uncover either overpriced or under-priced properties. From this, RealtyCheck can avoid the risk of being overcharged or reap benefits by quickly trading undervalued ones. Secondly, the automated data mining process enables RealtyCheck to make prompt data-driven decisions (Provost & Fawcett, 2013). Moreover, as up-to-date data are automatically provided to the system to perform such predictive analysis tasks, the company will likely achieve competitive advantages well ahead of time over its competitors (Provost & Fawcett, 2013). Thirdly, with the predictions, RealtyCheck can plan its marketing strategy according to future market trends and customer needs (Varma et al., 2018). Considering the specified business objectives, the system shall be judged as a successful project if profits from investing activities increase by 15% and the system is finished on time and within budget.

2.2 Data understanding

The second stage, understanding the data, focuses on identifying, collecting, and analyzing the relevant data sets (IBM, 2021). In detail, historical data is collected in most cases to perform predictions in the real estate sector (Gerek, 2014). Three relevant advantages of historical data to our system are insights into past events, events evolution, and forecasting events (Selim, 2009). However, as our targeted

variable - house prices- changes rapidly, historical data's prediction power and reliability fluctuate accordingly. To illustrate, property prices, according to the municipality of Rotterdam (2022), increased by approximately 103% in the last ten years and are continuously swinging every day (Statista, 2022). Further, historical data in the real estate sector vary in types, generally including commercial property data, transactional data, mortgage loan & lender data, pre-foreclosure data, ownership data, and commercial tenant data (Fuerst & Haddad, 2020). Meanwhile, sources to acquire real estate data are open market data, public records, brokerage reports, or consultancy reports (Fuerst & Haddad, 2020). Each data type and source have different characteristics and contribute differently to predicting house prices. Hence, adequate understandings of data exploration and quality verification are advisable for RealtyCheck to complete this data mining goal.

2.3 Data preparation

During the third stage of the CRISP-DM cycle, the data is prepared for developing the automated system. Based on the company's business objectives of making smarter real estate investments, we suggest RealtyCheck collect commercial property data from two sources: open market data and public records, specifically Rotterdam municipality's local property records. According to Gerek (2014), four essential metrics that help conduct proper real estate property valuation are locational, conditional, structural, and commercial. Open market data give insights into commercial aspects and promptly reflects the market trends. On the other hand, public records ensure reliable data about locational, conditional, and structural metrics (Pagourtzi et al., 2003; Gerek, 2014).

Locational metrics cover the characteristics of a property's specific location (Gerek, 2014). Distance to essential facilities such as schools, parks, or hospitals and to amenities such as highways, main roads, and centrum areas greatly determine the property's attractiveness (Aluko, 2011). Moreover, the prevalence of criminality has been observed to drastically impact house prices as it compromises buyer perception of safety (Kellekci & Berköz, 2006). Next, conditional metrics assess the house’s current state of affairs, which essentially involve the property's year of construction, subjective conditions, energy consumption level, and the type of permits. Over the past decades, construction standards have improved, focusing more on the high-energy efficiency structure of buildings, including insulation quality, heating installation, natural ventilation, solar systems, etc. (Brounen & Kok, 2011). A higher energy label implies a more sustainable and cost-saving energy consumption (Brounen & Kok, 2011), thereby increasing property value. The third metric, structural, as its name suggests, compose of asset type, architectural design, living area, number of (bed)rooms, and number of floors. These attributes are considered the most important valuations determinants (Selim, 2009). Finally, attributes in commercial metrics directly reveal the market valuations of a property and its potential profitability. For RealtyCheck, a real estate firm, this can assist them in making smarter investments due to more accurate predictions. Table 1 below summarizes the suggested metrics and the related attributes to be collected from the two mentioned sources.

Metrics	Attribute	Measurement levels
Locational	Local zoning type	Nominal
	Neighborhood safety	Ordinal
	Distance to essential facilities (e.g., hospitals, schools, , markets…)	Ratio
	Distance to amenities (e.g., local highways, , centrum, …)	Ratio
Conditional	Construction year of the property	Interval
	Energy label	Ordinal
	Overall condition (subjective)	Ordinal
	Building permits	Binominal
Structural	Asset type (house or apartment)	Binominal
	Total living area in square meters	Ratio
	Number of rooms	Interval
	Number of bedrooms	Interval
	Number of floors	Interval
	Architectural design	Nominal
Commercial	Potential renovation costs	Ratio
	Amount of collectible foreclosure	Ratio
	Mortgage loan	Ratio
	Last transaction amount	Ratio
	Market listing price	Ratio
	Number of days on the market	Interval

Table 1. Suggested metrics for property valuation

2.4 Modelling

2.4.1 Building the model

The fourth step in the CRISP-DM cycle is to build and assess different models based on various modeling techniques. As previously mentioned, historical commercial property data from open market data sources and public records are utilized to develop our model. As property prices are already defined as our target variable, the Supervised method is applied to the automated system. Moreover, concerning the business objective presented in the first step (section 1.1), Predictive modeling is chosen over Causal modeling. Predictive modeling is a supervised method that applies a statistical model or data mining algorithm to data to predict future observations (Belo, 2022). Of the two types of Predictive modeling,

Regression is the more appropriate model for this project goal, compared to Classification. To justify, Regression attempts to estimate or predict the numerical value of the target variable, in this case, house prices (Belo, 2022).

Regression modeling implements two types of formulas Mathematical (linear regression, logistic regression) or Rule-based (regression trees) (Belo, 2022). We incorporate Rule-based formula in the automated system for the following reasons. Our recommended datasets include variables that are measured on a nominal level (local zoning type and architectural design). Regression trees will have better average accuracy for such categorical independent variables than linear regression. Moreover, it is ambiguous whether the relationship between house prices and the mentioned independent variables is linear. If the relationship is non-linear, then a regression tree, by its nature, will probably fit the data better (Belo, 2022).

Nevertheless, certain assumptions are required when building the regression tree. In the first place, the collected training set is considered as one specific root which distinctively determines how a tree grows (Prasad, 2021). Next, each record is recursively distributed according to the attribute values (Prasad, 2021). And statistical approaches decide whether an attribute is a root or an internal node of the tree (Prasad, 2021). The final assumptions to be checked are normality and constant variance among the variables (Prasad, 2021). If these assumptions are met, the Regression tree is the suggested model to be developed for the system to predict property value.

2.4.2 Data collection frequency

Information collection frequency contributes to both the modelling and data preparation stages of the CRISP-DM cycle (IBM, 2021). The historical commercial property data from the two sources is stored under two prevalent formats: structured and unstructured.

On the one hand, in terms of structured data, the Rotterdam public municipality data is used to collect attributes related to the locational, conditional, and structural metrics. Most variables are entered once or updated in the database weekly to sync with new entries in the municipality database. Apart from open market data sources for commercial metrics, the data should be collected from the Top 5 banks in the Netherlands. If possible, the system may also use data purchased from 3rd party regarding the most clicked mortgage type on the bank web pages. These commercial attributes should be updated every six months to determine the best time to release a property on the market (Gigya, 2015). Besides, historical data on the target variable - housing prices should be collected from popular real estate, i.e., NVM, and public release data, i.e., *oecd.org*. Since housing prices on these sources are updated daily, the frequency of such collection will be daily as well.

On the other hand, unstructured data should also be collected to serve as a benchmark, which helps standardize categorical attributes into ordinal ones, thus, increasing the efficiency of our Regression modelling. Web scraping and text mining are suggested as the two methods to collect data from online sources. By connecting Google search and Twitter Streaming AP to our system, RealtyCheck can search for specific keywords such as “best architectural design” or “the safest neighbourhood in Rotterdam” to

retrieve public insights into consumer purchase intent (Zhou & Tong, 2022) (Google, 2022). Even though a Streaming API protocol continuously updates the unstructured data, its benchmark shall be updated at the same time with the evaluation step below

2.5 Evaluation

The fifth step in the CRISP-DM cycle assesses the performance and reviews the process of the model developed. As specified in the first step (section 2.1), the business success criterion is that RealtyCheck can gain an extra 15% of profits from their investing decisions by incorporating the automated system. Hence, the system result is evaluated on how accurately its built-in model predicts the price of a property compared to that property's marketing listing price. As for process review, we recommend agile methods. Following one of the agile methods, the scrum process, the system should be reviewed every two to four weeks (Scrum.org, 2020). In practice, a dedicated team shall inspect the system and communicate to product owners, who can decide on potential improvements in the data preparation and modelling steps. The product owner will then prioritize new features to the sprint cycle and deploy them to the system. Each update shall include complete documentation (Scrum.org, 2020).

2.6 Deployment

The sixth step, deployment, discusses the formal integration of the automated system to accomplish RealtyCheck business objectives (IBM, 2021). A successful deployment of this automated system requires that the right people receive the right information (IBM, 2021). Firstly, RealtyCheck's decision-making team should be informed of precisely predicted prices of attractive properties. Also, they need to be explained why this property is considered attractive and how attributes contributed to the price prediction. Secondly, the real estate agents will have to be updated with the new attractive property so they can perform onsite inspections and/ or negotiate better deals. Finally, database experts who monitor the real estate market and commercial property data should be revised with information usage and potential attributes to be included in the datasets in future projects. The step after planning for deployment and planning for monitoring and maintenance. Here, a similar scrum process in the evaluation step is adopted. Intentively, project-related members shall assess system performance every two to four weeks. A benefit of a scrum process is that new system features are demonstrated to RealtyCheck early and often. RealtyCheck can inspect progress and prioritize maintenance (Scrum.org, 2020). The final steps are producing a final report and conducting a final review of the automated system.

3. Conclusion

In brief, to accomplish RealtyCheck business objectives, the automated system shall identify attractive real estate properties and predicts their prices before market listing. We identified fifteen relevant attributes from four metrics: locational, conditional, structural, and commercial. And information to be collected is commercial property data from open market data and public records. Regression tree is the recommended model. And the model performance is assessed based on valuations prediction accuracy through the scum process.

Part II – Building a prediction model

1. Introduction

The second part of this report serves as an internal technical support document backing up the proposal to RealtyCheck outlined in the section above. The overall objective is to help the company predict which price range (out of five ranges) a real estate property belongs in. We begin by describing the training data set collected by our colleague. Then we explain our selection of the Gradient Boosted Tree Model, followed by an interpretation of this model's performance on testing data. Finally, we discuss our results, limitations, and the potential usefulness of our model to RealtyCheck's business model. We use Rapid Miner Studio throughout the whole model-building process, including selecting relevant variables and models, training the model, and optimizing parameters.

2. Training data description

The data collected for the training set were scraped from a real estate website that provides information about all properties available for sale in Rotterdam. The available information in our training set includes 4,888 records that belong to the following 32 attributes:

VARIABLE	ROLE	TYPE	DESCRIPTION
id	identifier	numeric	Unique property identifier
price_category	target	multinomial	Price range; 5 categories; see table above
year	predictor	numeric	Construction year
zipcode	predictor	multinomial	ZIP code of the property
total_rooms	predictor	numeric	Total number of rooms
bedrooms	predictor	numeric	Number of bedrooms
n_weeks_old	predictor	numeric	For how many weeks has the property been listed
living_area	predictor	numeric	Total living area (in square meters)
other_area	predictor	numeric	Other areas (in square meters)
total_area	predictor	numeric	Total area (in square meters)
monthly_contrib	predictor	numeric	Mandatory monthly contribution to owners association
n_photos	predictor	numeric	Number of photos posted
n_photos_360	predictor	numeric	Number of 3d photos
type_of_construction	predictor	multinomial	Type of construction
other_indoor_space	predictor	numeric	Other indoor areas (in square meters)
energy_label	predictor	multinomial	Energy label
located_on	predictor	numeric	Floor number
own_ground	predictor	binomial	Whether the property is located in own ground
flg_missing_year	predictor	binomial	Whether the variable year was missing
flg_missing_total_rooms	predictor	binomial	Whether the variable total_rooms was missing
flg_missing_bedrooms	predictor	binomial	Whether the variable bedrooms was missing
flg_missing_n_weeks_old	predictor	binomial	Whether the variable n_weeks_old was missing
flg_missing_living_area	predictor	binomial	Whether the variable living_area was missing
flg_missing_other_area	predictor	binomial	Whether the variable other_area was missing
flg_missing_total_area	predictor	binomial	Whether the variable total_area was missing
flg_missing_monthly_contrib	predictor	binomial	Whether the variable monthly_contrib was missing
flg_missing_n_photos	predictor	binomial	Whether the variable n_photos was missing
flg_missing_n_photos_360	predictor	binomial	Whether the variable n_photos_360 was missing
flg_missing_type_of_construction	predictor	binomial	Whether the variable type_of_construction was missing
flg_missing_other_indoor_space	predictor	binomial	Whether the variable other_indoor_space was missing
flg_missing_located_on	predictor	binomial	Whether the variable located_on was missing
flg_missing_own_ground	predictor	binomial	Whether the variable own_ground was missing

Table 2. Variables description

However, 7 attributes were omitted when building the model: flg_missing_livving_area, flg_missing_total_area, flg_missing_type_of_construction, flg_missing_year, total_area, type_of_construction, and id. The initial six attributes were not included due to high stability (>95%) while attribute id was left out as it offers no predictive insight. We retrieve this outcome from the Select inputs step in the Auto model extension in Rapid Miner.

3. Gradient Boosted Tree model

3.1 Model selection

The process of building and weighing different models was conducted in the Auto Model extension in Rapid Miner Studio. In detail, we first specify our problems in the prediction class. In the second step, we input the collected training data sets and define price_category as the targeted variable (label). Then we select the relevant columns to make predictions. In the dataset, seven variables are irrelevant, as discussed before. And in the last step, RapidMiner helps generates the results of different models, as shown in Table 3 below. By default, we were able to compare the results of Naive Bayes, Generalized Linear Model, Logistic Regression, Deep Learning, Decision Tree, Random Forest, and Gradient Boosted Trees (XGBoost). We prioritize accuracy, and Gradient Boosted Trees yield the highest accuracy performance at 61.9% with a standard deviation of 2.1%.

Model	Accuracy	Standard Deviation	Gains
Naive Bayes	39.4%	± 1.2%	432
Generalized Linear Model	47.5%	± 2.0%	652
Logistic Regression	42.4%	± 2.6%	506
Fast Large Margin	42.7%	± 1.4%	528
Deep Learning	58.1%	± 2.3%	944
Decision Tree	52.9%	± 0.5%	810
Random Forest	55.9%	± 1.6%	896
Gradient Boosted Trees	61.9%	± 2.1%	1,046

Table 3. Models results

3.2 Parameters Optimization

After selecting Gradient Boosted Trees as the preferred model, we performed additional parameters' optimization on the training set. In the Optimize Parameters (grid) process, we selected three parameters of the Gradient Boosted Trees to be optimized: min_rows (the minimum number of rows to assign to the terminal nodes.), maximal_depth (the tree depth), and learning_rateI (the learning rate). For min_rows, we set the range from 1 to 200, with 50 steps. Next, maximal_depth and learning rate were evaluated from 1 to 300 with 50 steps and from 0.01 to 0.99 with 30 steps, respectively. We applied a linear scale for all three processes. As a result, 15,751 parameter combinations were derived. The combination that yields the highest accuracy (0.7505) on the training set was *min_rows* = 20, *maximal_depth* = 41, and *learning_rate* = 0.15. Note that the Optimize Parameters (grid) process is combined with the Cross-validation process. And both processes are performed in RapidMiner Studio.

3.3 Performance assessment

As mentioned above, the accuracy of 0.7505 is the model performance on the training dataset. However, Kaggle is assessing the model performance based on its prediction accuracy on the testing data set, which consists of examples that the model has not seen before. In other words, the performance we

are presented with is the *generalization performance***.** Generalization performance, in this case, refers to the *Gradient Boosted Trees* model's ability to predict the listing price of a property from previously unseen data. The model generalization performance scored at 0.77250, implying that this model could correctly predict the listing price of a property with a 77.25 percent chance. The variables contributing to the model performance are discussed in more detail in the following paragraph.

3.4 Weight of variables importance

Table 4 below summarizes the variables' Weight by Information Gain in our Gradient Boosted Tree model. The values are ranked from high to low. The higher the weight of an attribute, the more relevant to the model. Moreover, the weights are normalized into the range from 0 to 1.

Attribute	Weight	Attribute	Weight
living_area	1.00000	other_indoor_space	0.05285
energy_label	0.58428	flg_missing_total_rooms	0.05214
year	0.42770	flg_missing_bedrooms	0.04829
total_rooms	0.41118	flg_missing_n_photos	0.04584
bedrooms	0.36780	n_photos_360	0.03649
zipcode	0.32859	flg_missing_n_photos_360	0.02079
n_photos	0.23940	flg_missing_other_indoor_space	0.01606
flg_missing_located_on	0.23310	n_weeks_old	0.01409
located_on	0.18644	flg_missing_own_ground	0.01010
monthly_contrib	0.10714	own_ground	0.00443
flg_missing_monthly_contrib	0.10714	flg_missing_n_weeks_old	0.00035
other_area	0.10127	flg_missing_other_area	0

Table 4. Normalized weights by information gain

Among the total 25 variables, the top 5 attributes with the highest weights are *living_area*, *energy_label*, *year*, total_rooms, and *bedrooms*, among which *living_area* significantly outperforms other attributes, weighted 1. Hence, variable living_area contributes the most for the model performance. The rest 4 variables are of high importance, as their weights are from 0.4 to around 0.6. The influence of variables such as *own_ground*, *n_weeks_old*, *n_photos_360* and *other_indoor_space* on the model performance is negligible compared with other non-binary variables.

Among the binary flag variables indicating whether the relevant variables are missing, only *flg_missing_located_on* is weighted 0.23310 suggesting that whether the information about the floor

number is missing can have a medium influence on the model performance, and the availability of other variables does not have a significant impact on the model performance.

4. Results discussion

Overall, the model performs on both the training and testing set, proved by the accuracy performance of 0.7505 and 0.7725, respectively. Regarding the model's accuracy on the testing set, the model can correctly predict 77.25% of the price range of a property given new data points. Therefore, RealtyCheck may use this model to predict the listing price of a property and anticipate fluctuations in the real estate market if the error rate of 22.75% is within its acceptance range. Furthermore, the flexibility in the Gradient Boosted Tree model allows RealtyCheck to make predictions from new attributes while maintaining accuracy at a certain level. By its nature, Gradient Boosted Tree provides several parameter options that are applicable to different scenarios. Also, it can work with various data types (University of Cincinnati, 2016). However, RealtyCheck should gather more training data and perform an extensive simulation to ensure the generalization of this model.

5. Limitations and assumptions

Certain limitations exist in our methodology. First, the used training dataset poses the following threats to validity. As we scraped secondary data from an online platform, we cannot verify the accuracy and quality of the data source. Moreover, the cleaned-up data may omit important records, thus, creating bias in the training dataset and in the model. Also, the data collected includes houses in different locations and may not correctly reflect the real estate prices in Rotterdam. Next, we did not examine the learning rate of the model on the training set. Hence, by gathering more observations, we could improve the performance of the model. Finally, we cannot assure extreme precision in the web scraping tool used for collecting the data - *webscraper.io*.

Second, the 31 chosen variables might not be optimal for this prediction task. Other variables that are also useful to determine the value of a property includes *Neighborhood safety, potential renovation cost, distance to essential facilities, number of floors,* and *energy labels*. Besides, as the targeted variable, house prices, changes rapidly, the prediction power of independent variables fluctuates accordingly. Hence, the reliability of this model has been affected accordingly.

Third, the chosen model, Gradient Boosted Trees, also has some limitations. We optimize the parameters through the automatic process in Rapid Miner Studio, which creates subprocesses for all combinations of selected values of the parameters and then delivers the optimal parameter values. So, the chosen parameters are biased toward the training dataset. Furthermore, the Gradient Boosted Trees model generates several weak models, which may result in overfitting (Park & Ho, 2021). And such ensemble modeling methods can be time-consuming and computer memory exhaustive (Jozefowicz et al, 2016).

Regarding the mentioned limitations, RealtyCheck should challenge the subsequent assumptions when applying this model. First, the collected secondary data used for training the model is valid, and its variables are useful in predicting house/apartment prices in Rotterdam. Second, the web scarpers extension: *webscarper.io* present no or very limited measurement error. Third, the prediction power of

chosen variables remains unchanged over time. And last, a general assumption when using a Gradient Boosted Tree is that encoded integer values for each input variable have an ordinal relationship.

6. Conclusion

In conclusion, the Gradient Boosted tree model is selected based on the secondary data collected due to its highest accuracy on the testing set. With the defined parameters, this model yields an accuracy performance of 0.7505 on the training set and 0.7725 on the testing set. In practice, RealtyCheck may use this model to predict the price range of property in Rotterdam with an accuracy of up to 77.25%.

References

Abdulhafedh, A. (2016). Crash frequency analysis. *Journal of Transportation Technologies*, *06*(04), 169–180. https://doi.org/10.4236/jtts.2016.64017

Aluko, O. (2011). The Effects of Location and Neighbourhood Attributes on Housing Values in Metropolitan Lagos. *Ethiopian Journal of Environmental Studies and Management*, *4*(2). https://doi.org/10.4314/ejesm.v4i2.8

Belo, R. (2022). *3.1—Introduction to Predictive Modeling—Annotated.pdf: Big Data Management and Analytics*. https://canvas.eur.nl/courses/40606/files/72464609?module_item_id=889665

Boelhouwer, P. (2017). The role of government and financial institutions during a housing market crisis: a case study of the Netherlands. International Journal of Housing Policy, 17(4), 591-602.

Braun, J., Burghof, H. P., Langer, J., & Einar Sommervoll, D. (2022). The Volatility of Housing Prices: Do Different Types of Financial Intermediaries Affect Housing Market Cycles Differently? *The Journal of Real Estate Finance and Economics*, 1-32.

Brounen, D., & Kok, N. (2011). On the economics of energy labels in the housing market. *Journal of Environmental Economics and Management*, *62*(2), 166–179. https://doi.org/10.1016/j.jeem.2010.11.006

Chakure, A. (2022, February 10). *What Is Decision Tree Classification?* Built In. Retrieved 27 November 2022, from https://builtin.com/data-science/classification-tree

Fernando, J. (2022, November 3). *R-squared formula, regression, and interpretations*. Investopedia. Retrieved November 30, 2022, from https://www.investopedia.com/terms/r/r-squared.asp#:~:text=In%20finance%2C%20an%20R%2DSquared,depend%20on%20the%20specific%20analysis.

Gerek, I. H. (2014). House selling price assessment using two different adaptive neuro-fuzzy techniques. *Automation in Construction*, *41*, 33–39. https://doi.org/10.1016/j.autcon.2014.02.002

Gigya. (2015). *Uncovering the hidden costs of third-party data - IBM*. IBM. Retrieved November 30, 2022, from https://www-50.ibm.com/partnerworld/gsd/showimage.do?id=40879

Google. (2022). *Google APIs Explorer*. Google apis explorer | google developers. Retrieved November 30, 2022, from https://developers.google.com/apis-explorer

Hayes, A. (2022, June 24). *Multiple Linear Regression (MLR) Definition, Formula, and Example.* Investopedia. Retrieved 30 November 2022, from https://www.investopedia.com/terms/m/mlr.asp

Hotz, B. N. (2022, November 13). *What is CRISP DM?* Data Science Process Alliance. Retrieved 24 November 2022, from https://www.datascience-pm.com/crisp-dm-2/

IBM. (2021, August 17). *CRISP-DM Help Overview*. Retrieved 24 November 2022, from https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). *Exploring the Limits of Language Modeling.* doi:10.48550/ARXIV.1602.02410

Kellekci, M. L., & Berköz, L. (2006). Mass Housing: User Satisfaction in Housing and its Environment in Istanbul, Turkey. *European Journal of Housing Policy, 6*(1), 77–99. https://doi.org/10.1080/14616710600587654

Mahmood, M. S. (2022, July 24). *Assumptions of Multiple Linear Regression - Towards Data Science.* Medium. Retrieved 30 November 2022, from https://towardsdatascience.com/assumptions-of-multiple-linear-regression-d16f2eb8a2e7

Marill, K. A. (2004). Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression. *Academic Emergency Medicine, 11*(1), 94–102. https://doi.org/10.1111/j.1553-2712.2004.tb01379.x

McMillen, D. P., & Thorsnes, P. (2006). Housing Renovations and the Quantile Repeat-Sales Price Index. *Real Estate Economics, 34*(4), 567–584. https://doi.org/10.1111/j.1540-6229.2006.00179.x

Mohamed, I., Jusoh, Y., Abdullah, R., & Nor, R. N. H. (2019, July). *Measuring the performance of Big Data Analytics Process - ResearchGate*. ResearchGate. Retrieved November 28, 2022, from https://www.researchgate.net/publication/342887515_MEASURING_THE_PERFORMANCE_OF_BIG_DATA_ANALYTICS_PROCESS

Park, Y., & Ho, J. C. (2021). Tackling Overfitting in Boosting for Noisy Healthcare Data. *IEEE Transactions on Knowledge and Data Engineering*, 2995-3006. doi:10.1109/TKDE.2019.2959988

Petchko, K. (2018). Data and Methodology. *How to Write About Economics and Public Policy*, 241–270. https://doi.org/10.1016/b978-0-12-813010-0.00013-2

Provost, F. and Fawcett, T., 2013, *Data Science for Business: What you need to know about data mining and data-analytic thinking*, O'Reilly Media

SAS. (n.d.). *Predictive Analytics: What it is and why it matters.* Retrieved 30 November 2022, from https://www.sas.com/en_us/insights/analytics/predictive-analytics.html

Scrum.org. (n.d.). *What is Scrum?* Scrum.org. Retrieved November 30 2022, from https://www.scrum.org/resources/what-is-scrum

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*(2), 2843–2852. https://doi.org/10.1016/j.eswa.2008.01.044

Sharma, M. (2018, December 20). *Multinomial Logistic Regression Using R.* Data Science Beginners. Retrieved 27 November 2022, from https://datasciencebeginners.com/2018/12/20/multinomial-logistic-regression-using-r/

Statista. (2022, 22 november). *Average purchase price of residential property in the Netherlands 1995-2021*. Statista. Retrieved 27 November 2022, from https://www.statista.com/statistics/593642/ average-purchase-price-of-dwellings-in-the-netherlands/

University of Cincinnati. (2016). *Gradient Boosting Machines*. Retrieved November 28 2022, from http://uc-r.github.io/gbm_regression

Varma, A., Sarma, A., Doshi, S & Nair, R (2018). House Price Prediction Using Machine Learning and Neural Networks. *Second International Conference on Inventive Communication and Computational Technologies*, 1936-1939. 10.1109/ICICCT.2018.8473231

Yinger, J. (1979). Estimating the relationship between location and the price of housing. *Journal of Regional Science, 19*(3), 271–286. https://doi.org/10.1111/j.1467-9787.1979.tb00594.x

Zhou, R., & Tong, L. (2022, January 1). *A study on the influencing factors of consumers' purchase intention during Livestreaming E-Commerce: The mediating effect of emotion*. Frontiers. Retrieved November 30, 2022, from https://doi.org/10.3389/fpsyg.2022.903023

Zietz, J., Zietz, E. N., & Sirmans, G. S. (2007). Determinants of House Prices: A Quantile Regression Approach. *The Journal of Real Estate Finance and Economics, 37*(4), 317–333. https://doi.org/10.1007/s11146-007-9053-7