Creating accurate valuation models for real estate properties

WISHAL M SRI RANGAN, University of Twente, The Netherlands

Housing is a necessity for society to function both economically and socially. Knowing the accurate value of a property is a major asset for both buyers and sellers. Buyers will be able to distinguish between good and bad deals and also be able to negotiate the price of a property during a transaction. Similarly, the seller will be able to know the exact market value of their home before listing it for sale. Automated valuation models are computer programs that make use of real estate property information to predict the value for a property. The use of automated valuation models serves as a tool for both buyers and sellers to appraise property. In this research, I will be evaluating different machine learning models as well as exploring the impact of property features in predicting real estate property prices. This research will be using a data set which includes a wide list of selling prices and a list of features for real estate properties in Cyprus. This study seeks to identify the machine learning approach that yields the most accurate feature-based prediction models, with a strong emphasis on data pre-processing.

Additional Key Words and Phrases: valuation models, property prices, real estate, machine learning, non-linear models

1 INTRODUCTION

Housing is an integral component of economic and social development and has a profound impact on achieving economic growth. It is considered a necessity, along with food and medical care, and can play a significant role in developing primary and secondary financial markets. Additionally, housing is an effective way to build wealth, as it appreciates in value and provides secure premises for income-generating activities [7]. These socioeconomic benefits mean housing prices can have a direct effect on quality of life, as well as a city's economic development and sustainability [14].

Accurate property valuations necessitate fair real estate market values, although manual real estate valuation is time consuming and may not consider all the factors that affect pricing. The hedonic pricing framework, a linear regression model, was used to estimate real estate property prices based on various characteristics [12]. However, this approach is not equipped to handle non-linear and unstructured data such as text. Research, however, has demonstrated that non-linear machine learning models can generate much more precise property valuations than hedonic regression models, as Baur explains in her paper [3].

1.1 Problem Statement

The use of machine learning has been crucial in the improvement of the accurate valuation of real estate properties. Artificial intelligence when applied to automated valuation of real estate properties has strong potential to provide accurate property price predictions and with the right amount of useful data can be more accurate than manual appraisals. Machine learning techniques have the advantage of finding patterns in high dimensional data which leads to prediction results that are difficult to beat by traditional regression approaches. It is clear that as automated valuation models become more advanced their accuracy and usability will increase with it [5]. In this paper, we will examine the different machine learning algorithms and the property attributes that can be included in constructing an accurate model to predict property prices in Cyprus. The methodology section of this paper will present the data collection, data pre-processing, machine learning models, hyperparameter tuning and evaluation metrics used in this research. Additionally, the results of this study will have the ability to pinpoint which features are significant in determining the value of a real estate property.

1.2 Research Questions

The main question this paper aims to answer is:

Can we increase the precision of ML-based valuation models when feeding them with additional real estate contextual information?

This question is best answered when split into two questions, each covering an aspect of building an valuation model.

- **RQ1** : How to select the right attributes and accordingly pre-process them?
- **RQ2** : Which of the selected machine learning models yields the most accurate prediction on real estate property valuation?

At the conclusion of our research, we should be able to determine the most accurate machine learning model for assessing properties based on the chosen evaluation metrics, as well as recognize which property attributes have a significant impact on a property's value.

1.3 Related Work

In the 2022 paper, Soltani [13] trains four machine learning models with three decades worth of housing price data from metropolitian Adelaide making use of multiple linear regression (MLR), decision tree (DT), random forest (RF) and gradient-boosted tree (GBT) and uses R-squared (R^2), root mean square error (RMSE) and mean absolute error (MAE) as evaluation metrics. The findings from this study showed that GBT and RF generate better performance compared to the other machine learning models. The research also showed that the relationship between housing price and features to be non-linear and therefore non-linear based regression models like DT have better performance that the MLR. Furthermore, it was found that GBT and RF yielded better predictive performance. In this study the spatiotemporal lag (ST-lag) variable was explored to see if it improved the predictive accuracy of the models and based on the findings, it was found that the ST-lag variable significantly improves the accuracy of machine learning based property price prediction models.

Similarly in 2021, Dieudonné et al. [15] reviews real estate price estimation in France. They compare seven machine learning techniques but the best predictors for each evaluation metric were found

TScIT 38, February 3, 2023, Enschede, The Netherlands

to be the artificial neural networks (ANN), RF and the k-nearest neighbors(KNN). To assess the predictive performance, the research makes use of the measures, mean absolute error (MAE), root mean square error (RMSE) and mean squared logarithmic error (MSLE) among several others. Furthermore, this research explores the impact of geocoding as one of the independent variables in the models. When compared with the experiment without geocoding, ensemble learning algorithms such as RF, GBT and adaptive boost (adaboost) out-performed all other algorithms for each metric. In summary, the real estate property value predictions with geocoding were better than predictions without geocoding.

Lasota et al. [8] in 2013 performed an investigation of property valuation models based on decision tree ensembles with five different methods where they concluded that the rotation forest (RTF) provided the best performance. This study is backed by another research conducted by Rodríguez et al. [11] where it was found that RTF has similar diversity-accuracy pattern to Bagging, but is slightly more accurate and diverse which resulted in statistically huge differences in favor of Rotation Forest. This model will be further explored in my research.

In Soltani's research [13], it is found that GBT and RF produced the best predictive performance out of the four machine learning methods that were used. Dieudonné [15] concluded that the best predictors in their experiment with geocoding were found to be RF, GBT and adaboost, the three models being ensemble learning algorithms. Moreover, Baur concludes in her paper [3] that the best to worst performing models were GBT, RF and support vector regression (SVM) respectively.

The chosen features in each of the related research were collected and combined as seen in Table 2. This table is categorized into three main categories: accessibility, structural and neighbourhood features. By combining the features from all the works, processing of the data can be done more efficiently and a deeper understanding of all the factors on what affects a house price can be made. From further investigation into socio-economic features, it was found that the house price index (HPI), per capita personal income among other neighbourhood features listed in Table 2 had an impact on property prices as mentioned by Pan et al. [9].

In Table 1, the evaluation metric scores for each of the models selected in the research conducted by Soltani [13] and Dieudonné [15] can be found. T

2 METHODOLOGY AND APPROACH

This section will cover the steps for performing this research. This includes model selection, hyperparameter tuning, evaluation metric selection and data collection.

2.1 Data Collection

The real estate dataset that will be used for this paper has been collected from a client in Cyprus which contains a list of selling prices of the properties along with another list of features for each property.

The original data set requires data pre-processing due to a lot of missing entries as well as missing features that would have to be added through the help of other services. This will be covered in the next section. Moreover, the features can be categorised into three categories namely accessibility, structural and neighbourhood. Accessibility refers to features such as number of hospitals, schools and parks in proximity to a property. The category structural refers to features of a property such as the number of floors or the land area of the property. Neighbourhood refers to socio-economic features such as the population density and unemployment. A summary of the chosen features can be seen in Table 3.

2.2 Data Pre-processing

Prior to training machine learning algorithms, data pre-processing must be done. This includes cleaning the data, normalizing the data, and selecting features. The steps are involved are as follows:

(1) Merging all the data files and sheets into a singular data set utilizing the property id as a key.
(2) Removing columns that do not contain any property information. This includes columns that are duplicated or contain property comments that do not have any valuable information.
(3) Removing duplicate entries with the same property id that was created as a result of the merge.
(4) Extracting 'Share Numerator' and 'Share Denominator' from the column 'Share' in order to realise the actual price of a property. This was done since the data set contained properties that had shared ownership. In short, shared ownership allows for a percentage of a property to be bought. In order to train the valuation models with accurate results, it was necessary to provide the full value of a property.
(5) Creating a new column 'Residential Type' using several other column in order to determine if a property was an apartment or a house. Ordinal encoding was used to 'Residential Type'. Similarly a new column 'Parking' was also generated using the same approach. One-Hot Encoding was used for this field as it had low cardinality.
(6) Creating a new column 'Grade' which graded properties based on how many properties were sold from the same town and district in Cyprus on a scale of 1-10.
(7) Since the original data set did not contain a number of accessibility variables such as number of schools or hospitals, this data was collected with the aid of a web service that uses 7 variables namely 'District', 'Quarter', 'Block', 'Village', 'Sheet', 'Plan' and 'Parcel' and returns the distance to each nearest amenity from the property within a 50 km radius. This was done with the aid of Google's Places API [1] that returns features in the accessibility category such as police stations, schools and hospitals. The search radius of 50km was limited by the API which only allows a maximum of 50,000 meters. Moreover, additional features that deal with features that lie in the neighbourhood category namely population density, unemployed population and purchasing power was found

[1]The Places API is a service that returns information about places using HTTP requests. Places are defined within this API as establishments, geographic locations, or prominent points of interest.

		Evaluation Metrics					
Paper	Models	R^2	RMSE	MAE	MSLE	Q1	MedAE
Soltani	DT	**-ST lag 0.579 +ST lag 0.797**	**-ST lag 0.135 +ST lag 0.094**	**-ST lag 0.106 +ST lag 0.064**	-	-	-
	RF	**-ST lag 0.662 +ST lag 0.875**	**-ST lag 0.109 +ST lag 0.087**	**-ST lag 0.086 +ST lag 0.059**	-	-	-
	GBT	**-ST lag 0.674 +ST lag 0.896**	**-ST lag 0.101 +ST lag 0.086**	**-ST lag 0.084 +ST lag 0.058**	-	-	-
Dieudonné	ANN	**0.64**	**82138**	**52687**	**0.14**	15121	33178
	RF	**0.74**	**71150**	**44786**	**0.11**	11123	26878
	adaBoost	**0.74**	**70923**	**45238**	**0.12**	11968	29059
	GBT	**0.72**	**73242**	**45272**	**0.11**	10090	26286
	Linear Regression	**0.48**	**97628**	**65472**	**0.3**	21524	45454
	SVM	**0.32**	**109840**	**71448**	**0.23**	22790	47983

Table 1. Evaluation metrics and models used from related work

using ArcGis Online [2]. The full list of these features can be seen in Table 3.

(8) Due to this limitation of the search radius, some of the entries in the data set were missing some feature values in the accessibility category. In order to resolve this, imputation processing was performed. A better alternative to removing rows and columns with missing values is to impute averages to the missing entries. This helps prevent useful data that would have been lost if discarded.

(9) After collecting all the required features, the next step is to normalize the data. The features area, population density, purchasing power and unemployed population were normalized data in order to improve the training stability and performance of the models.

(10) Finally, the Inter-Quartile Range method (IQR) is used to remove outlier data points on the columns. Equation 1 refers to the lower bound and Equation 2 refers to the upper bound. The entries in the data set with values smaller than the lower bound and values larger than the upper bound are excluded for each column. The IQR method was applied to four features area, purchasing power, unemployed population and population density. As seen in figure

$$x < Q_i - 1.5 \times IQR \tag{1}$$

$$Q_3 + 1.5 - IQR < x \tag{2}$$

Following data pre-processing, the final data set was comprised of 1481 entries and 29 features. The initial data set was made up of 137000 entries and 42 features, including duplicate rows. A selection of the features is shown in the correlation matrix depicted in Fig. 1. Examination of the correlation matrix reveals that most of the features have a correlation coefficient of 1. However, the features Purchasing Power and Purchasing Power: Index shows the existence of multicollinearity due to their coefficients being 1 and therefore Purchasing Power: Index will be removed from the data set for model training.

3 RESULTS

3.1 Valuation Models

Based on these prior studies and work, I have narrowed down to four machine learning methods that I plan to apply to my research.

- **Random Forest (RF)** : A supervised machine learning algorithm introduced by Breiman [4] that is based on multiple decision trees (DTs) trained with randomly generated subsets of the data set and finally combined into a singular model to make accurate predictions and also preventing the risk of over fitting as explained in Soltani's paper [13]. The RF algorithm can detect non-linear relationships between the dependent and independent variables chosen.
- **Gradient-Boosted Tree (GBT)** : Similar to RF, GBT is an ensemble machine learning model based on DTs. When compared to RF, GBTs train each DT one at a time where as RF trains several DTs simultaneously. GBTs trains each DT iteratively to minimize a loss function. This loss function on every iteration gets lowered by the GBT while running on the training data.
- **Adaptive Boosting (adaboost)** : A learning technique that aims to increase the efficiency of a learning system by boosting. The basic principle in boosting is when a singular strong learner gets compromised by multiple weak learners. This results in greater stability than a single complex tree. [15]
- **Rotation Forest (RTF)** : - Rodriguez et al. [11] created this ensemble method in 2006. The RTF algorithm randomly rotating the feature space and training a different classifier for each rotation, it combines the advantages of random subspace and bagging approaches. A group of classifiers are produced as a result, and they are then integrated to reach a classification decision. RTF performs better than existing ensemble methods, like random forests. Although initially proposed for classification, Pardo et.al [10] showed that adaption for regression was possible through their study obtaining favourable results compared to AdaBoost and Random Forest.

3.2 Hyperparameter Tuning

Each of the four machine learning algorithms have a set of hyperparameters. Table 5 shows the parameters that were chosen for this

[2]ArcGIS Online is a cloud-based mapping and analysis solution. It is a tool used to make maps and analyse data

Category	Feature	Type (unit)	Mentioned in
Accessibility	Euclidian distance to public transport	double	Soltani
	Network distance to nearest airport	double	Soltani
	Euclidian distance to beach	double	Soltani
	Network distance to primary school	double	Soltani
	Distance to nearest roadside significant sites	double	Soltani
	Network distance to nearest secondary school	double	Soltani
	Network distance to nearest university	double	Soltani
	Network distance to nearest train line stops	double	Soltani
	Network distance to nearest tram line stops	double	Soltani
	Network distance to nearest hospital	double	Soltani
	Network distance to nearest public space	double	Soltani
	Network distance to nearest main shopping center	double	Soltani, Lasota
	Network distance to nearest entertainment center	double	Soltani
	Network distance to city center	double	Lasota
Structural	The year of construction	int (year)	Soltani, Lasota
	No. of ensuite rooms	int (count)	Soltani
	Land area	double (sq meter)	Soltani, Dieudonné
	Floor area	double (sq meter)	Soltani, Dieudonné, Lasota
	No. of floors	int (count)	Soltani, Lasota, Baur
	No. of bedrooms	int (count)	Soltani, Lasota, Baur
	No. of bathrooms	int (count)	-
	Sewerage availability	boolean	Soltani
	Water availability	boolean	Soltani
	Ownership type of dwelling	str (categorical variable)	Soltani
	Property category	str (categorical variable)	Soltani, Dieudonné
	Roof material type	str (categorical variable)	Soltani
	Wall material type	str (categorical variable)	Soltani
	Building style	str (categorical variable)	Soltani
	Street number	int	Dieudonné
	Street type	str	Dieudonné
	Street name	str	Dieudonné
	Postal code	str	Dieudonné
	City	str	Dieudonné
Neighbourhood	Population per area	double	Soltani
	Population change in the research period	double	Soltani
	Spatiotemporal lag	double	Soltani
	House price index (HPI)	double	Pan
	Per capita personal income	double	Pan
	Labor force	double	Pan
	Unemployment rate	double	Pan
	Per capita personal income growth	double	Pan
	Growth rate of labor force	double	Pan
	Nominal mortgage rate	double	Pan
	Real mortgage rate	double	Pan
	Inflation rate	double	Pan
	HPI percentage changes	double	Pan
	Equity ratio	double	Pan
	Cost-income ratio	double	Pan

Table 2. Features from all related work

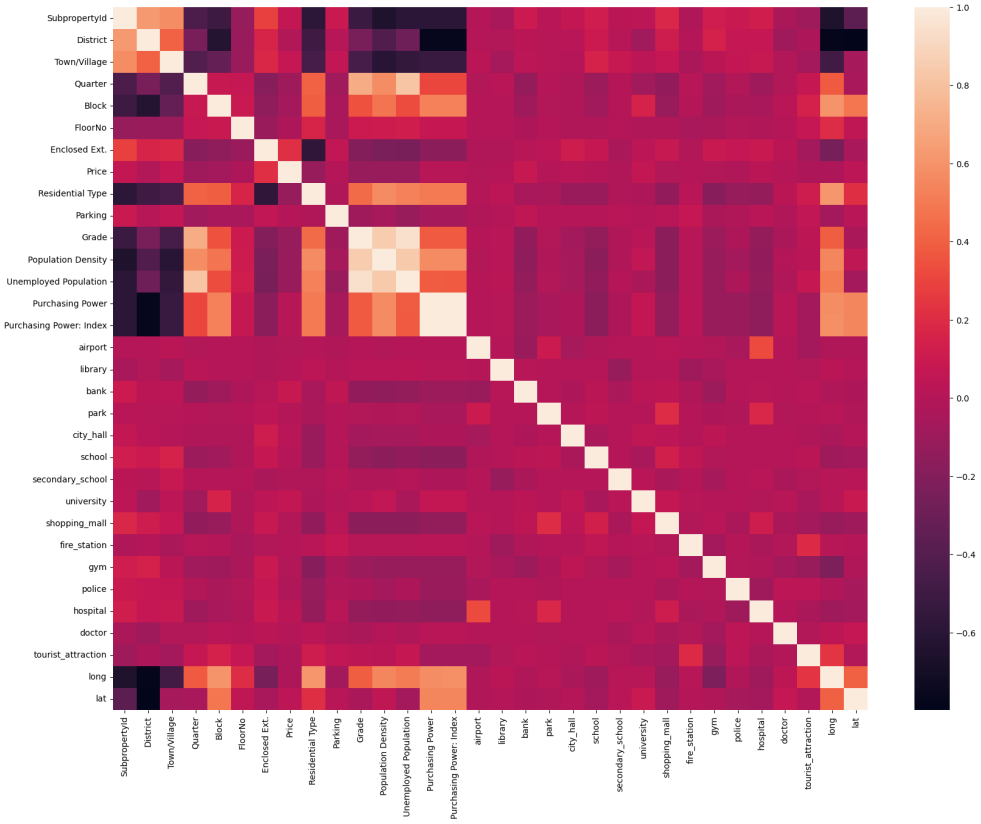

Fig. 1. Correlation matrix between a few features selected

research. Grid search and random search are the two most used parameter selection functions. Grid search works by creating a grid of parameter combinations, trains the model for each combination, and finally selects the combination that yields the best results. Random search chooses a value for each hyperparameter based on random sampling. In order to choose the best possible parameters for this research both random search and grid search will be used. Random search will be used to identify the general range of values for the parameters to be chosen and grid search will be used to target the most ideal parameter values.

3.3 Evaluation Metrics

The evaluation metrics that will be used to evaluate the models is an important measure in order to see how accurate the models are in predicting the property prices. To evaluate the models, I have decided to incorporate four statistical measures as follows:

- **R-Squared**(R^2) :

$$1 - \frac{\sum_1^n (y_i - \widehat{y_i})^2}{\sum_1^n (y_i - \overline{y_i})^2} \tag{3}$$

- **Root Mean Square Error (RMSE)**

$$\sqrt{\frac{1}{n}\sum_1^n (y_i - \widehat{y_i})^2} \tag{4}$$

- **Mean Absolute Error (MAE)**

$$\frac{1}{n}\sum_1^n \|\widehat{y_i} - y_i\| \tag{5}$$

- **Mean Squared Logarithmic Error (MSLE)**

$$\frac{1}{n}\sum_1^n (\log(y_i + 1) - \log(\widehat{y_i} + 1))^2 \tag{6}$$

Where n is the number of data points, $\widehat{y_i}$ is the predicted value, y_i is the actual value and $\overline{y_i}$ is the mean value.

3.4 Machine learning model results

After performing data pre-processing, the refined data set is ready to be trained by the four machine learning algorithms that have been chosen. As mentioned, the hyperparameters were chosen with the help of using random Search and grid search. The following subsections covers the source of these algorithms.

Category	Feature	Type (unit)	Explanation
Accessibility	Airport	double	Distance to nearest airport
	Library	double	Distance to nearest library
	Bank	double	Distance to nearest bank
	Park	double	Distance to nearest park
	City Hall	double	Distance to nearest city hall
	School	double	Distance to nearest school
	Secondary School	double	Distance to nearest secondary school
	University	double	Distance to nearest university
	Shopping Mall	double	Distance to nearest shopping mall
	Fire Station	double	Distance to nearest fire station
	Gym	double	Distance to nearest gym
	Police	double	Distance to nearest police station
	Hospital	double	Distance to nearest hospital
	Doctor	double	Distance to nearest doctor's office
	Tourist Attraction	double	Distance to nearest tourist attraction
	Supermarket	double	Distance to nearest supermarket
	Parking	boolean	Availability of parking
Structural	District	int	Ordinal encoding used for each district
	Town	int	Ordinal encoding used for each town
	Quarter	int	Ordinal encoding used for each quarter
	Block	int	Ordinal encoding used for each block
	No. of floors	int (count)	Number of floors in the property
	Area	double	In square meters of a property's total area
	Latitude	double	Latitude coordinate of the property
	Longitude	double	Longitude coordinate of the property
	Residential Type	boolean	(1) refers to an apartment and (0) refers to a house.
Neighbourhood	Population Density	double	Population density of the district the property lies in
	Unemployed Population	double	Unemployed population of the district the property lies in
	Purchasing Power	double	Purchasing power of the district the property lies in
	Grade	int (ordinal grades)	Explained in paper

Table 3. Description of the chosen features.

Model	R2	RMSE	MAE	MSLE
Random Forest	0.6708	74331.9899	47094.3335	0.1330
Gradient-Boosted Tree	0.5756	84402.4863	53949.8003	0.1898
Adaptive Boosting	0.7771	61172.8420	36898.4869	0.0729
Rotation Forest	0.6326	78536.8005	45837.8670	0.1080

Table 4. Results of the machine learning algorithms

Valuation Model	Hyperparameters
Random Forest	number of estimators, minimum samples split, minimum sample leaf, maximum features, maximum depth
Gradient-Boosted Tree	learning rates, number of estimators, minimum samples split, minimum sample leaf, maximum features, maximum depth
Adaptive Boosting	type of estimator, decision tree max depth, number of estimators, learning rate, loss function
Rotation Forest	number of estimators, minimum samples split, minimum sample leaf, maximum features, maximum depth

Table 5. Hyperparameters chosen for the machine learning algorithms

3.4.1 Random Forest algorithm. The RF algorithm implemented was taken from scikit-learn [3] and the chosen hyperparameters can be found in Table 5.

3.4.2 Gradient-Boosted algorithm. The GBT algorithm implemented was also taken from scikit-learn and the chosen hyperparameters can be found in Table 5.

3.4.3 Adaptive Boosting algorithm. The adaBoost algorithm implemented was also taken from scikit-learn and the chosen hyperparameters can be found in Table 5. The estimator used was a decision tree regressor.

3.4.4 Rotation Forest algorithm. The Rotation Forest algorithm [4] was based on scikit-learn's random forest module and was implemented using the algorithm developed by Rodríguez et al. [11]. The rotation forest algorithm shares the same hyperparameters as the random forest algorithm since it is derived from the later.

[3] https://scikit-learn.org/stable/index.html
[4] https://github.com/digital-idiot/RotationForest

Table 4 shows the resulting metrics R2, RMSE, MAE and MSLE for each of the selected machine learning models. The RMSE and MAE scores represents the error in the models in Euros (€).

3.4.5 Best Performing Model. The RF model using 5-fold cross validation and random search along with grid search resulted in a R2 score of 0.6708 and a MSLE score of 0.1330. Next the GBT model with the same pre training procedure as the RF model, resulted in a R2 score of 0.5756. The GBT model provided a MSLE score of 0.1330. The adaboost algorithm scores 0.7771 on the R2 metric and 0.0729 on the MSLE metric. Finally, the RTF algorithm resulted in scores 0.6326 and 0.1080 for the R2 and MSLE metric. Using the test set, the results show that adaboost performance better than the three other models with a R2 of 0.7771 and MSLE of 0.1080. Similarly the adaboost model scores RMSE and MAE estimated to be 78536 and 45837. The GBT model returned the lowest R2 score implying the data set has a bad fit with the model. This may be due to the fact that Gradient-Boosted trees use shallow trees that underfit. The chosen max depth of the trees can be found in Appendix A and Table 6 which explains the reasoning behind the low R2 score for GBT compared to the other models. In comparison to performance metrics from the studied related work [13] [15] found in Table 1, the models RF, adaboost and RTF along with the data set chosen for this research returned similar results for all metrics R2, RMSE, MAE and MSLE. However, in Soltani's research, the spatiotemporal lag variable was also explored as a feature and the incorporation of it resulted in better evaluation metric scores when compared to the results of this research using the Cyprus data set.

3.4.6 Feature Importance. Since Adaptive Boosting had the best performance out of the other models, the feature importance was retrieved from it. As seen in Figure 2, the results show that purchasing power, number of floors of the property, area of the property, distance to amenities such as shopping malls, universities, schools, banks and police stations as well as the latitude and longitude coordinates are the most important features in determining a property's value. Purchasing power of the district the property lies in seems to have the largest impact on a property's price based on the Adaptive Boosting model.

Additionally, Figure 2 shows that neighbourhood attributes such as Purchasing Power and Population Density can have a profound impact on a property's valuation compared to structural and accessibility features. However it also revealed that that some of the features included for model training had no impact whatsoever on property valuation. The results from the feature importance figure helps answers the research question RQ1 as it helped better understanding which attributes have an impact in a valuation model.

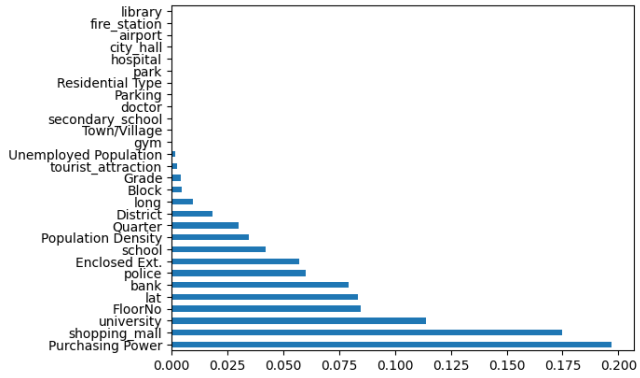

Fig. 2. Feature importance chart of all the features selected using Adaptive Boosting

4 CONCLUSION

This research aims to enhance the understanding of property valuation by considering all possible features. The study also aims to investigate the effectiveness of various machine learning models in predicting property prices. Four machine learning models, including Random Forest, Gradient-Boosted Tree, Adaptive Boosting and Rotation Forest were trained using property price data from Cyprus. Results from the research indicate that machine learning algorithms can effectively predict house prices. Evaluation metrics such as R2, RMSE, MAE, and MSLE were used to compare the performance of the four models and it was found that the Adaptive Boosting model resulted in the best performance metric scores. Initially it was expected that the Rotation Forest model would result in the best performance due to prior research regarding the experiments conducted by Bagnall et al. [2] which showed that on average, rotation forest is better than common ensemble alternatives such as Random Forest, Gradient Boosted Tree and Adaptive Boosting. However it is clear that all the four ensemble machine learning models can result in favourable performances. It can be concluded that Adaptive Boosting resulted in the best predictive performance out of the four models, answering research question RQ2.

This research has the potential to benefit property owners and buyers by providing a more accurate way to appraise property based on various features that may not be considered without the use of machine learning algorithms. Furthermore, it suggests that certain machine learning models may be better suited for predicting property prices and highlights the need for further research in this area. The importance of socio-economic features namely purchasing power and population density were notable qualities found from the results of this research. It calls for further research into social and economic affairs of a property's proximity having an impact on its valuation since in general automated valuation models were more focused on structural and accessibility features [3][15][8][6].

5 FUTURE WORK

In regards to future work, there are a lot of additional features that could have been included into this data set and the only limiting factor was the number of features that could have been collected in this research. Table 2 shows all the possible features that could be incorporated to improve the performance of the models. As seen from the results of the feature importance in Figure 2, socio-economic features have a strong impact on property valuation and therefore if more features as listed in the Neighbourhood category in Table 2 such as house price index (HPI), labor force and inflation rate were explored, the valuation models may predict better resulting in better performance metrics. The machine learning models and the

hyperparameters used could have be further optimized with similar methods such as random Search and grid search. Another limiting factor for this research was the time constraint which prevented further investigation into more ideal hyperparameters due to the exhausting processing times of both random Search and grid search.

Moreover, the current data set had a lot of entries with missing values for certain features which reduced the total size of the data that the machine learning models could be trained on. This constraint could have been avoided if some other algorithms such as the XGBoost algorithm and the K-nearest neighbour algorithm(KNN) was used as part of this research. According to a research conducted by Aulia et al. [1], the XGBoost algorithm can handle missing values without the need imputation pre-processing, a method that was employed for my research. The results in their paper show that the XGBoost model without any imputation pre-processing has a similar accuracy to the XGBoost models with imputation pre-processing. In another research involving real estate property valuation by Karshiev et al. [6], the KNN-based most correlated features (KNN–MCF) algorithm was used to deal with the missing data entries. The results of the method employed showed better performance with the KNN-MCF algorithm. For this research, imputation pre-processing was used although it may have resulted in inaccurate entries especially since it was used to fill in missing values in accessibility features such as distance to amenities.

In addition to the above, future work could also include collecting data from other regions or countries to expand the scope of the study. The research could be extended by including other types of data such as demographic, economic and social data which would provide a more comprehensive understanding of the factors that influence property prices. Additionally, more advanced machine learning algorithms such as deep learning models could be used to improve the accuracy of the predictions. Overall, there are many opportunities for further research in this area and the results of this study serve as a starting point for further future work.

ACKNOWLEDGMENTS

I would like to thank my supervisor Andreas Kamilaris and research associate Asfa Jamil for their help in collecting and providing the data set required for this project as well as guiding me through the process of developing the machine learning models.

REFERENCES

[1] Deandra Aulia and Hendri Murfi. "XGBoost in handling missing values for life insurance risk prediction". In: *SN Applied Sciences* 2 (Aug. 2020). DOI: 10.1007/s42452-020-3128-y.

[2] Anthony Bagnall et al. "Is rotation forest the best classifier for problems with continuous features?" In: (Sept. 2018).

[3] Katharina Baur. "Automated Real Estate Valuation with Machine Learning Models Using Property Descriptions". In: *Expert Systems with Applications* 213 (2023). DOI: https://doi.org/10.1016/j.eswa.2022.119147.

[4] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001). DOI: https://doi.org/10.1023/a:1010933404324.

[5] B. A. J. Hilgers. "Automated Valuation Models for Commercial Real Estate in the Netherlands: Traditional Regression versus Machine Learning Techniques". In: (Oct. 2018).

[6] Sanjar Karshiev et al. "Missing Data Imputation for Geolocation-based Price Prediction Using KNN–MCF Method". In: *ISPRS International Journal of Geo-Information* 9 (Apr. 2020), p. 227. DOI: 10.3390/ijgi9040227.

[7] Duane Kissick. "Housing for All: Essential to Economic, Social, and Civic Development". In: (2006).

Hyperparameters	Value
Random Forest	
n_estimators	120
min_samples_split	12
min_samples_leaf	3
max_features	3
max_depth	80
Gradient-Boosted Tree	
learning_rate	0.1
n_estimators	94
min_samples_split	0.1
min_samples_leaf	0.1
max_features	None
max_depth	8
Adaptive Boosting	
base_estimator	DecisionTreeRegressor
max_depth	11
n_estimators	10
learning_rate	1.5
loss	'exponential'
Rotation Forest	
n_estimators	200
min_samples_split	10
min_samples_leaf	4
max_features	'sqrt'
max_depth	178

Table 6. Optimal hyperparameters for the Cyprus real estate property data set

[8] Tadeusz Lasota. "Investigation of Property Valuation Models Based on Decision Tree Ensembles Built over Noised Data". In: *Computational Collective Intelligence. Technologies and Applications* (2013). DOI: https://doi.org/10.1007/978-3-642-40495-5_42.

[9] Huiran Pan and Chun Wang. "House prices, bank instability, and economic growth: Evidence from the threshold model". In: *Journal of Banking Finance* 37.5 (2013), pp. 1720–1732. ISSN: 0378-4266. DOI: https://doi.org/10.1016/j.jbankfin.2013.01.018. URL: https://www.sciencedirect.com/science/article/pii/S0378426613000435.

[10] Carlos Pardo et al. "Rotation Forests for regression". In: *Applied Mathematics and Computation* 219.19 (2013), pp. 9914–9924. ISSN: 0096-3003. DOI: https://doi.org/10.1016/j.amc.2013.03.139. URL: https://www.sciencedirect.com/science/article/pii/S0096300313004128.

[11] J.J Rodriguez. "Rotation Forest: A New Classifier Ensemble Method". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.10 (2006). DOI: https://doi.org/10.1109/tpami.2006.211.

[12] Sherwin Rosen. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition". In: *ournal of Political Economy* 82.1 (1974). DOI: https://doi.org/10.1086/260169.

[13] Ali Soltani. "Housing Price Prediction Incorporating Spatio-Temporal Dependency into Machine Learning Algorithms". In: *Cities* 131 (2022). DOI: https://doi.org/10.1016/j.cities.2022.103941.

[14] D Streimikiene. "Quality of Life and Housing". In: *International Journal of Information and Education Technology* 5.2 (2015). DOI: https://doi.org/10.7763/ijiet.2015.v5.491.

[15] Serge Nyawa Tchuente Dieudonné. "Real Estate Price Estimation in French Cities Using Geocoding and Machine Learning". In: *Annals of Operations Research* 308.1-2 (2021). DOI: https://doi.org/10.1007/s10479-021-03932-5.

A APPENDIX A

As mentioned in the methodology, the hyperparameters were chosen by firstly using random search to find a general search point and then grid search to get better performance out of the models. Both the parameter selection functions used 5-fold cross validation to evaluate the models. The optimal hyperparameters that could be found during the span of this research is found in Table 6.