Ankara Real Estate Price Forecasting Using Machine Learning Techniques

Oguzhan Ergun
Computer Engineering
Middle East Technical University NCC
Guzelyurt, TRNC
e215191@metu.edu.tr

Abstract— **Housing price forecasting is an utmost important matter not only for the property owners but also for the buyers. Various machine learning techniques are applied to analyze the real-estate transactional data across the world to contrive useful models. This paper covers extracting and deriving features from a real estate ad dataset, which contributes to the perceived value of the property and apply different regression techniques with those new features to obtain better performing models in terms of root mean squared error for predicting the estate prices in Ankara, Turkey.**

I. Introduction

Apart from being a reflection of a countries economic state, the housing market is also a key factor for forming the direction of the economy. Housing construction supports the economy by raising house sales, contracting, and investments. It also boosts demand for other related sectors, such as construction goods and household goods. (Li et al., 2011). Housing price forecasting may be used for estimating the demand for those new projects or for determining whether it will be a lucrative investment to make. Real-estate market forecasting studies try to reveal patterns in growth and analyze the perceived value of houses. In recent years, the improvement in accessibility of machine learning techniques and the widespread availability of the data have made real estate markets a feasible field for studies using machine learning algorithms.

Nowadays, it is possible to find a range of research that analyzes the real estate market using machine learning methods. However, research which emphasize on feature extraction and generation using internal and external data sets and later applying machine learning techniques to forecast housing price are in the minority.

The purpose of this study is to extract features using the real estate ad data set, explore useful connections between this data set, and a social media activity data set, try to develop valuable models for predicting the housing price market in Ankara, Turkey. Successful models may help both parties in their decision-making process. Moreover, revealing essential features that affect the perceived value of the properties may be useful for construction companies and real estate agencies.

Following parts of the paper structured in a way such that Section 2 conducts a literature review on housing market prediction using different machine learning algorithms. Section 3 includes exploration of the dataset, identification of data quality issues, explanation of applied pre-processing steps to increase the data quality. Section 4 provides a detailed description of machine learning techniques used in the study. Section 5 presents and discusses the results of the different models, and Section 6 summarizes the study.

II. Literature Review

House price valuation studies focus on the estimation of house values using machine learning techniques [1], [2]. These studies attempt to develop useful models to forecast the real estate price; their training set has variables such as location, land size, and the number of rooms. They are using Support Vector Regression (SVR) and its combination with other techniques for house value prediction. For example, Chen et al. [2] combine SVR and Stepwise to project house prices effectively. Moreover, Principle component analysis is also used in the pre-processing and model development with SVR. Besides, the joint usage of Stepwise and SVR is commonly used for Dimension Reduction of High-Dimensional Datasets, as mentioned by Chou et al. [3]. In another paper written by Nissan et al. [9], the real estate property prices in Montreal are analyzed, and regression models have been developed. The information on the real estate listings was scraped from Canadian real estate ad platforms. They have predicted both asking and sold prices of the real estate assets based on the features same as the first two papers. Moreover, they had applied feature engineering and derived new features such as the nearest police station and fire station, which were extracted from the Montreal Open Data Portal, their Random Forest Regression model, obtained meager error value for the target feature “sold price” [9].

In another paper by Yu et al. [4], which was about real estate price prediction, various machine learning techniques were used, such as for continuous property prices Lasso, Ridge, SVR, and Random Forest regression. For the individual price ranges, they had predicted with classification methods, including Naive Bayes, logistic regression, SVM classification, and Random Forest classification [4]. This paper allows readers to gain extensive knowledge about the difference of the methods. Moreover, it gives insight into a proper model selection based on the structure of the data set.

Ng et al. [10] have developed various models to predict house prices in London. Regression methods such as linear, Bayesian linear, relevance vector machines and gaussian process were used. Furthermore, results of the predictions were demonstrated in a mobile application using a heat map. Gaussian processes for regression was used as their final model due to its flexibility and probabilistic approach to learning and model selection. Similar to ensemble learning method in order to process the large data they have divided it and developed local models. Final predictions were the result of these local models’ outputs.

Many more researches were conducted along this line with different regression techniques. Selim [11] investigated the determinants of real estate price in Turkey by using the hedonic model and illustrated that artificial neural network could be a better approach for prediction of the house price in Turkey. Park and Bae [12] developed a general prediction model based on machine learning methods such as C4.5, RIPPER, Naive Bayesian, and AdaBoost and compared their performance for the classification accuracy. Wang et al. [7] proposed a unique model based on SVM to predict the average house price in different years, meanwhile, the authors concluded that PSO algorithm could effectively find out the parameters of SVM.

III. Data Exploration and pre-processing

A. Datasets

Three datasets were used in this study. All of them were downloaded from Kaggle website [8]. Data set which contains real estate ad training data has 40,670 observations and 13 variables. Data set which contains real estate ad test data has 4,524 observations and 13 variables same as the training set. Each observation represents a real estate ad published on a turkish real estate ad platform between June 2019 and July 2019. It is possible to arrange variables in three sections.

- Ad related variables include Ad id, Ad title, Posted, Currency, Price .
- Property related variables include M^2,Type
- Location-based variables include Latitude, Longitude, District, Quarter name, City, Population

The target variable is the price, which is a numeric type variable.

Last data set contains twitter data which has 2-3% of the shared tweets with Twitter's Streaming API. It has only the tweets which have geotagged in the Ankara region. It has 15,265 observations and 8 variables. It is possible to create 2 groups for this variables.

- Location-based variables include Latitude, Longitude, Quarter name
- Tweet related variables include Tweet, User id, Tweet Date and X which represents index number for the row.

Quarter name variable will be mainly used to extract new features in the following steps.

B. Pre-processing

Initially, it is vital to have an insight into data and to obtain such insight data must be prepared first. Pre-processing started with inspection for the missing variables in the data sets. Fortunately, data sets had no entries with missing data.

However, despite having a unique ad identification number, multiple listings for the same property were detected, those duplicates were deleted from the training set but not touched in the test set as the challenge required exactly 4,524 observation predictions.

Later on, an anomaly detected on the M^2 (area) variable in the training set. Although the M^2 variable was supposed to be numerical, it was categorical, row causing that issue detected and removed from the data set, and afterward, M^2 variable converted from categorical to numerical data.

Some ad-related variables pruned from the data set as they had no beneficial use case, those variables were,

- Ad id which has meaning inside the real estate ad platform
- Currency which was a variable having the same value for each observation
- City which was a variable having the same value for each observation
- Posted which was indicating the post date of the ad

Furthermore, outliers were also discovered and addressed. Outliers are defined as an observation which seems to be inconsistent with the remainder of the dataset [6].

Majority of the outliers were detected in variables, M^2 and price, instead of deleting them, pattern extraction was performed in the ad title variable which helped to correct most of them.

For example, if the actual value of M^2 variable was mentioned in the ad title variable for that outlier observation, it was transferred to the M^2 variable. Rest of the outliers which could not be fixed were pruned.

Finally, there was inconsistency in the type variable, which indicates the property's type. They were corrected using the same pattern extraction strategy and some factor levels which had same characteristics were merged.

After all cleaning procedures, the final data set without feature engineering had 36,160 observations and 9 variables. Type feature will be one-hot encoded and will be used in model development. Moreover, log transformation will be applied as the data is not balanced.

TABLE I. Real Estate ad data feautures

Name	Type	Description
Ad title	Nominal	Title of the ad
Latitude	Numerical	Latitude value of the property location
Longitude	Numerical	Longitude value of the property location
Type	Categorical	Type of the property
M^2	Numerical	Area value of the property (square meter)
District	Categorical	District name which property is registered based on location
Quarter name	Categorical	Quarter name which property is registered based on location
Quarter population	Numerical	Population of the registered Quarter
Price	Numerical	Property price (Target)

Categorical data had following number of levels after pre-processing.

- Type: 7 Levels
- District: 17 Levels
- Quarter name: 367 Levels

Tweet data set had no issues regarding to missing values. As for pre-processing some features were pruned as they had no use case for this study, those features were,

- X which representing index value for tweet
- User id which was showing twitter user id that send the tweet

- Tweet which had the content of tweet
- Tweet date which was demonstrating the tweet post time
- Tweet id which was representing id of the tweet

Resulting data set had 15,265 observations with 3 variables.

TABLE II. TWEET DATA FEAUTURES

Name	Type	Description
Latitude	Numerical	Latitude value of the tweet's geolocation
Longitude	Numerical	Longitude value of the tweet's geolocation
Quarter name	Categorical	Quarter name which imputed based on tweet location

C. Feature Engineering

As could be observed from table 1, the real estate data set provides only the fundamental features for the advertised property. Enriching the features would be beneficial in terms of increasing the model performance.

Previously pattern extraction or regular expression detection using the ad title feature was performed to fix outliers and the incorrect data. It was also possible to derive number of rooms available for certain types of property. Those type levels were "Daire", "Villa" and "Residence".

Table 3 demonstrates example for each type level that used to infer number of rooms feature.

TABLE III. AD TITLE FEATURE EXAMPLE

Ad title	Type Level	Pattern	Layout
Yapracık Toki 14. Bölge 3+1 137 M2	Daire	\\d+\\s*\\+\\s*\\d+	3+1
Bağlıca Süper Lüks 5+2 Villa	Villa	\\d+\\s*\\+\\s*\\d+	5+2
PORTAKAL ÇİÇEĞİ RESİDENCE 1+1 BALKONLU	Residence	\\d+\\s*\\+\\s*\\d+	1+1

Table 4 represents the performance of this approach when it was performed for both data set.

TABLE IV. NUMBER OF ROOMS EXTRACTION STATISTICS

Type	Set	Number of successful extractions	Number of total observations	Percentage
Daire	Training	23,114	30,328	76 %
	Test	2,741	3,701	74 %
Villa	Training	409	1,196	34 %
	Test	59	169	35 %
Residence	Training	485	601	81 %
	Test	45	62	72 %

Afterwards, random forest method was used with the other features to predict the number of rooms feature for the observations which did not have this information in their ad title feature. In the following sections details will be discussed.

Population feature of the data set was providing the population with respect to quarter, population with respect to district was derived by grouping the quarters by their district. Table 4 demonstrates the result of this operation.

TABLE V. DISTRICT POPULATION

District	Population
Polatlı	194,930
Gölbaşı	84,505
Çubuk	63,799
Kazan	28,852
Akyurt	17,961
Elmadağ	17,001
Beypazarı	12,996
Haymana	1,410
Ayaş	349

District	Population
Çankaya	887,915
Keçiören	828,166
Yenimahalle	587,559
Sincan	454,256
Mamak	452,227
Etimesgut	413,492
Altındağ	269,058
Pursaklar	215,015

Using the tweet data set, it was possible to derive quarter popularity and their district's popularity as well. In this context popularity indicates the frequency of the tweets that shared from the location.

Result of the following operation is shared in table 6.

TABLE VI. DISTRICT POPULARITY TABLE

District	Popularity
Çankaya	10,476
Keçiören	1,664
Yenimahalle	1,459
Etimesgut	832
Gölbaşı	363
Sincan	194
Polatlı	171
Mamak	101
Pursaklar	4
Çubuk	1

As a result, data set was enriched with new features, table 7 summarizes the added features.

TABLE VII. ADDED FEATURES

Name	Type	Description
Room size	Categorical	Layout of the property
District population	Numerical	Population of the district using sum of quarter populations
Quarter popularity	Numerical	Popularity of the quarter using tweet data set
District popularity	Numerical	Popularity of the district using sum of quarter popularities

D. Descriptive exploration

Aim of this section is to present the significant findings. Descriptive analysis such as correlation between variables and summary of the data set is given in the appendix section.

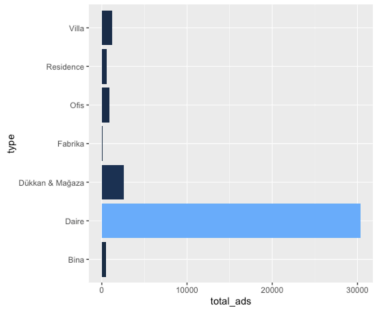


Fig. 1. Number of ads with respect to their property type

Figure 1 represents total number of ads given to the real estate platform with respect to property type, type level "Daire" was the most popular level among all. Whereas ads for type level "Fabrika" was quite rare.

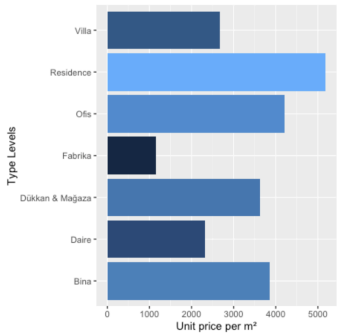


Fig. 2. Unit price per unit area values with respect to property types

Figure 2 demonstrates how unit price per M^2 varies across different type levels. For example, type level "Residence" had the highest price per are unit whereas type "Fabrika" had the lowest price per unit.

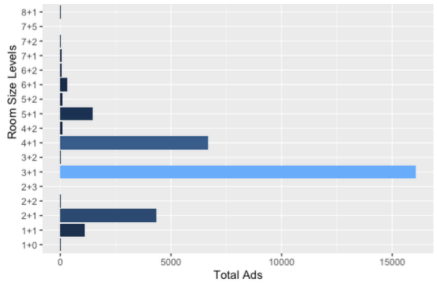


Fig. 3. Room size levels frequency

Figure 3 shows that the majority of the property having type "Daire" had room size feature with "3+1" layout. Followed by "2+1" and "1+1".

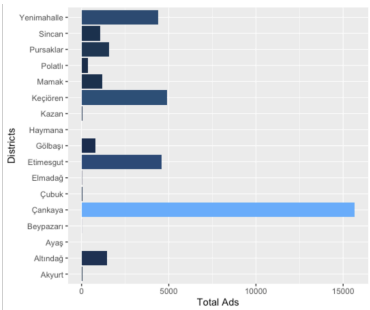


Fig. 4. District ad frequency

Figure 4 gives information about location of the properties that were listed in the real estate ad platform. District "Çankaya" had the highest frequency 15,661. "Ayaş", "Akyurt, "Beypazarı", "Çubuk", "Elmadağ", "Haymana" and "Kazan" districts had very low listings.

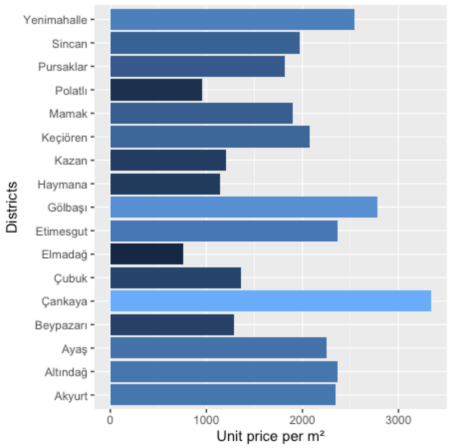


Fig. 5. Unit price per m^2 for districts

Figure 5 gives information about the most expensive district in terms of unit price per unit area. District "Çankaya" was the most expensive among all.

IV. Methodology

Model selection and evaluation

Apart from PCA, the stepwise technique will be used to determine the optimal number of features and later be supplied to various models using different techniques such as Linear Regression, Polynomial Regression, Lasso Regression, and XG Boost Regression. The primary model for evaluation will be based on Linear Regression; the performance measure will be the adjusted root mean squared error, according to MSE is the most popular tool to measure the quality of fit [6].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \quad (1)$$

Formula takes the actual value subtract it from the prediction and gets square of the difference which makes the value positive and then takes average of the result.

A. Principle Component Analysis

Principal component analysis (PCA), is used to reduce the complexity of the model. The complexity of the model is formed by the number of features in the model. For example, a model trained using three features is more complex than a model trained using two features [6].

Principal component analysis tries to reduce model complexity by identifying the essential features that explain the majority of data set variety. It is often used as a dimension reduction technique. It works with numerical data only. CATPCA which is a type of PCA could be used if the one wants to analyze categorical data as well.

TABLE VIII. PCA Results

Feature	Proportion of Variance
M^2	32.21%
Quarter population	22.45%
Quarter popularity	14.36%

Feature	Proportion of Variance
District popularition	12.40%
District popularity	10.54%
Latitude	5.35%
Longitude	2.69%

Table 8 demonstrates results of the PCA for the data set. Apart from "Latitude" and "Longitude" features, other features play an important role in the variance of the data set. It is possible to obtain 91.97% of the variance in the data set using 5 features. Thus, for the model formulation top 5 features will be used.

B. *Stepwise Regression*

Stepwise Regression constructs different regression models iteratively. The aim is to find the model with optimal number of features and least adjusted mean square error. Stepwise Regression also performs cross validation such as using k-folds to obtain proper results. Researcher may perform forward stepwise regression which includes more features in every iteration or backwards method which starts with all features and decrease the feature size every iteration, moreover; it is possible to perform a bi-directional training.

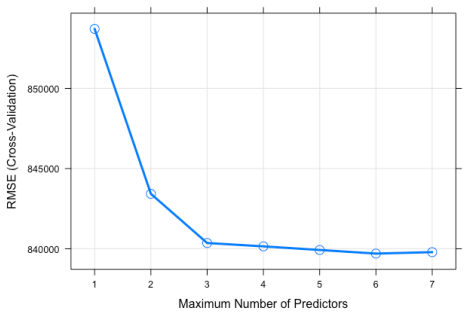


Fig. 6. Stepwise Regression results

Figure 6 demonstrates result of the stepwise regression using the numerical features that was also used in PCA analysis. In order to get a decent model, it is sufficient to have five numeric predictors as the RMSE tends to saturare, although more predictors could be used, redundant complexity should be avoided.

C. *Decision Trees*

Decision Trees are mostly used in classification, it is possible to use decision trees for regression as well. However, in this study decision trees are the building blocks of random forest method. Number of decision trees were used to determine the derived room size feature from the data set.

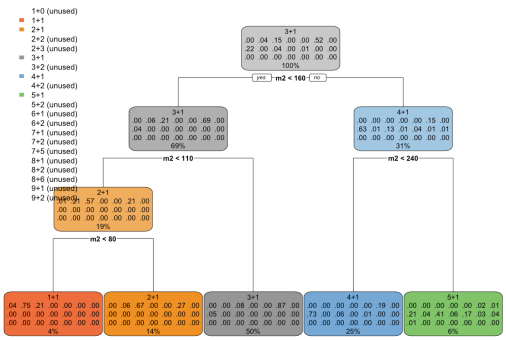


Fig. 7. Decision Tree for detecting room size

Figure 7 demonstrates how the decision-making process works for the given observation.

D. *Linear Regression*

Linear regression is widely used to model linear correlations, moreover; linear regression results faster than other regression techniques which do not have O(n) complexity. It uses training data features to predict the target feature. Those predictor's individual contributions are represented as coefficients of the linear formula. Once the formula is constructed, the model is fitted, one can get the predicted value just by plug-in the predictor values.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad (2)$$

Above formula represents a linear regression model.

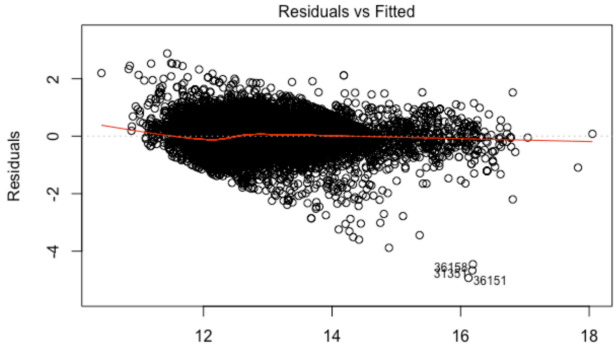


Fig. 8. Residuals vs fitted graph for linear regression model for the training set

Figure 8 shows residuals vs fitted graph for the linear model, this graph is useful for detecting non-linearity, unequal error variances and outliers.

E. *Polynomial Regression*

Polynomial regression is just a different version of linear regression. Complexity of the model is determined by the predictor feature having highest rank, thus the performance may vary. It tries to establish a polynomial formula between the target feature and the predictor features. Similar to linear regression once the coefficients for the predictors are found, model is fitted, predicted value could be obtained by plug-in predictor values. Below formula represents a polynomial regression model.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d + \epsilon_i \qquad (3)$$

F. *Random Forest*

Decision trees are the basic foundation of random forest method. The main idea of random forest method is to combine variety of decision trees and take average of them which results in overall unbiased, not overfitted model than decision tree models. Random Forest does not require cross validation for determining overfitting.

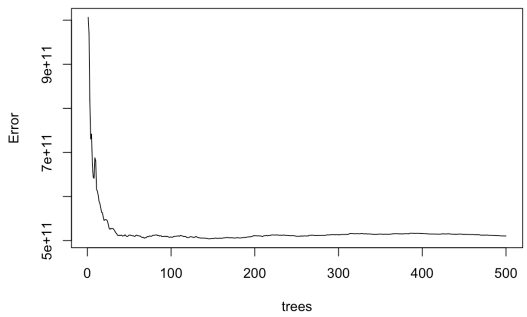


Fig. 9. Random forest error over number of trees in the forest

Figure 9 illustrates error rate over number of trees used to create a random forest for the training data set, as could be seen from the graph error rate decreases drastically between interval 0 to 100.

G. Support Vector Regression

Support Vector Regression (SVR) uses support vectors to construct a hyperplane which is going to be use for data segregation. Each observation in the set may be assumed as it is representing its own dimension. Support vectors are obtained by evaluating the chosen kernel across the training set observations and test set observations, outcome also represents the location of the test point in that dimension. Once the vector is obtained, it could be used to for regression tasks. Support vector regression is actually is alike linear regression but using countless dimensions.

Various kernels exist in SVM, however, there are four essential ones sigmoid, linear, radial basis function and polynomial. In this study radial basis function kernel was applied as the number of features was relatively less. Radial basis function is a well-suited kernel for such scenario. [5]

Optimized parameters for radial basis function kernel was determined using "tune.svm" function with cross validation, table 9 shows those parameter values.

TABLE IX. SVR MODEL PARAMETERS

SVR Model Parameters	
Cost	1
Gamma	0.0769
Epsilon	0.1
Number of support vectors	10,507

V. RESULTS / FINDINGS / COMPARISON

In this study R language was used for development and evaluation of the models. Development environment was RStudio on a macOS operating system.

First of all, contribution of derived features on the root mean squared error will be discussed for each regression technique. Afterwards, different model performances will be compared, as mentioned in the previous sections linear regression model will be the reference model and the other models' performance will be quantified based on the ratio to the linear regression model's root mean squared value.

Then, hybrid model (ensemble) approach will be discussed; training set will be divided based on the type feature of the observations. Since the majority of the observations consisted of having type level "Daire"," Villa" and "Residence" for this kind of observations, models with different formulas will be developed. Another reason for this approach was to make use of the derived room size feature which was applicable only to those levels.

In the end, predictions for those levels will be merged with the prediction made by other models for other type levels and performance of the hybrid model will be evaluated and compared. Contribution of popularity future which was extracted using tweets data set for linear regression model is demonstrated in the below table.

TABLE X. CONTRIBUTION OF POPULARITY FEATURE TO ENTIRE DATASET

Regression Method	Model Formula	RMSE
Linear	M^2 + Type + Q. population + D. population	784,800
	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity	780,500
Polynomial	M^2 + Type + Q. population + D. population	861,700
	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity	860,500
SVR	M^2 + Type + Q. population + D. population	776,406
	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity	762,605
Random Forest	M^2 + Type + Q. population + D. population	693,204
	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity	672,401

On 36,159 degrees of freedom Linear regression with derived features performed 0.05% better, 0.014% for polynomial, 2% better for SVR and the biggest jump observed was in random forest which was 3.1%, also it was the best performing model.

Another derived feature was room size. This feature was only available for a subset of type variable; "Daire", "Villa" and "Residence". Contribution of this feature is represented in the following table.

TABLE XI. CONTRIBUTION OF ROOM SIZE FEATURE TO SUBSET OF TRAINING SET

Regression Method	Model Formula	RMSE
Linear	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity	174,400
	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity + Room Size	170,100
Polynomial	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity	191,500
	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity + Room Size	179,200
SVR	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity	155,568
	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity + Room Size	153,300
Random Forest	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity	162,440
	M^2 + Type + Q. population + D. population + Q. popularity + D. popularity + Room Size	155,517

On 32,124 degrees of freedom, As could be observed, linear method performance increase by 2.3%, for the polynomial modal it was resulted in 6.9% increase, for the SVR it resulted in 1.5% increase. Finally, for the random forest 4.45%

increase was observed. Although the highest gain was in polynomial model, SVR performed best among them.

Results show that feature engineering contributes to performance of the regression model no matter which technique is used.

Finally, in the following table different regression models will be compared. All of the models are using the same number of features for the development.

Formula: $M^2 + Type + Q.\ population + D.\ population + Q.\ popularity$

TABLE XII. MODEL COMPARISION

Model	RMSE
Linear Regression	780,500
Polynomial Regression	860,500
Support Vector Regression	762,605
Random Forest Regression	672,401

On 36,159 degrees of freedom, As can be seen from the table Support Vector Regression performed 2.35% better than Linear Regression model whereas Polynomial Regression performed 2.31% worse than Linear Regression model. Random forest model performed 16% better which also made it the best performing model as well.

Moreover, in order to increase the performance hybrid model (ensemble learning) experiments were made. Main idea was to use additional feature which was available for a subset of the data set. Below formula was used to train models for the observations in the training set with type levels "Daire", "Villa", "Residence".

Model 1 Formula: $M^2 + Room\ Size + Type + Q.\ population + D.\ population + Q.\ popularity + D.\ popularity$
Model 2 Formula: $M^2 + Type + Q.\ population + D.\ population + Q.\ popularity + D.\ popularity$

TABLE XIII. ENSEMBLE MODEL COMPARISION

Ensemble Model (Model 1&2)	RMSE
Linear Regression	777,017
Polynomial Regression	775,060
Support Vector Regression	752,035
Random Forest Regression	654,336

On 36,159 degrees of freedom, Table 13 shows that how ensemble model helped decreasing the RMSE further, for example linear regression had 780,500 RMSE before and 777,017 after which means 0.045% performance gain. For polynomial results were 11% increase in performance, for Support Vector Regression it resulted in were 1.4% increase and finally for random forest, it resulted in 2.7% improvement over standard model depicted in table 12. Highest performing mode was random forest model and it was 19% better than the ensemble linear regression model and 19.3% better than standard linear model.

VI. CONCLUSION / DISCUSSION

In summary, this paper tries to find useful models for house price forecasting. It also exhibits the overall picture of the real estate ad market in Ankara. Initially, dataset is analyzed and prepared. Then, in the descriptive exploration section paper aims to find trends or preferences in the local market, it focuses on them and tries to derive additional features. Moreover, feature extraction from the training data with the usage of regular expressions were performed. Also, an external data set was used to derive new features. Random forest technique was not only used in prediction but was also used to determine observations without having information related to those derived (room size) features. Then, data transformation and reduction were applied by using Stepwise and PCA techniques. Various methods are then implemented such as linear regression, polynomial regression, support vector regression and random forest. Contribution of the new features as well as the models using different regression techniques were evaluated. It is seen that new features improved performance for all of the studied techniques.

Hybrid approach which is also known as Ensemble Learning model is used. Experiments shown that this approach provides performance increase over standard model prediction for any technique used in the study. Ensemble Learning method with support vector regression models and random forest models shown that it is a competitive approach. Thus, it could be used in other data sets with having similar structure. Finally, it is possible to use the same approach for different regions as the model is not using quarter name or district feature which is special just for a certain location (city).

REFERENCES

[1] Mu, J., Wu, F. & Zhang, A., 2014. Housing Value Forecasting Based on Machine Learning Methods. Abstract and Applied Analysis, (2014), p7

[2] Chen, J.-H. et al., 2017. Forecasting spatial dynamics of the housing market using Support Vector Machine. International Journal of Strategic Property Management, 21(3), pp.273–283.

[3] Chou, E. P., & Ko, T. W. (2017). Dimension Reduction of High-Dimensional Datasets Based on Stepwise SVM

[4] Yu, Jiafu Wu. 2016, "Real Estate Price Prediction with Regression and Classification" Yu, Jiafu Wu.

[5] Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

[6] James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112). New York: springer.

[7] X. Wang, J. Wen, Y. Zhang, and Y. Wang, "Real estate price forecasting based on svm optimized by pso," Optik-International Journal for Light and Electron Optics, vol. 125, no. 3, pp. 1439–1443, 2014.

[8] https://www.kaggle.com/c/metu-ncc-cng514-challenge/data

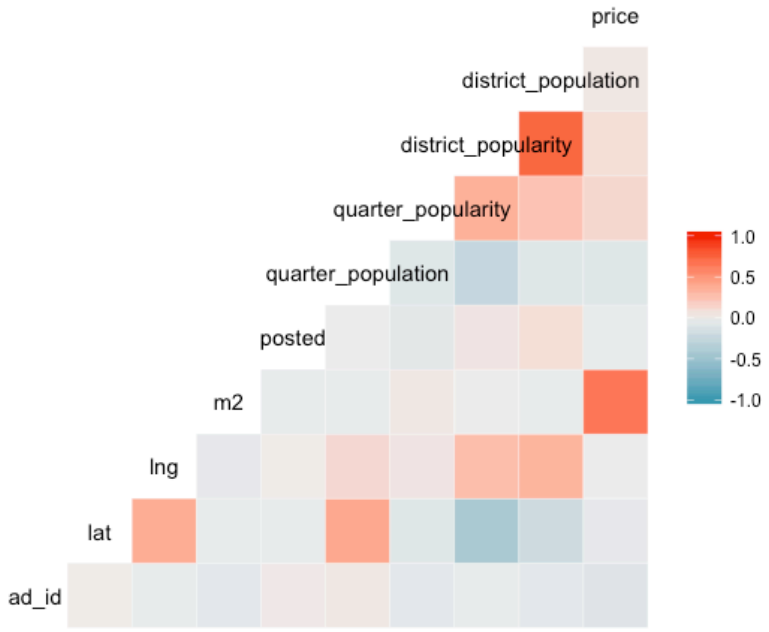
[9] Nissan Pow, Emil Janulewicz, Liu (Dave) Liu, 2016, "Applied Machine Learning Project 4 Prediction of real estate property prices in Montreal."

[10] Ng, A., & Deisenroth, M. (2015). Machine learning for a London housing price prediction mobile application. Technical Report, June 2015, Imperial College, London, UK.

[11] H. Selim, "Determinants of house prices in turkey: Hedonic regression versus artificial neural network," Expert Systems with Applications, vol. 36, no. 2, pp. 2843–2852, 2009.

[12] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data," Expert Syst. Appl., vol. 42, no. 6, pp. 2928–2934, 2015.

Appendix



Numeric value correlation graph

```
    price                     type           m2          quarter_popularity quarter_population
Min.   :   50000   Bina           :  516   Min.   :   11.0   Min.   :   0.00    Min.   :    67
1st Qu.:  205000   Daire          :30328   1st Qu.:  112.0   1st Qu.:   0.00    1st Qu.:  6969
Median :  310000   Dükkan & Mağaza: 2566   Median :  135.0   Median :   5.00    Median : 12821
Mean   :  505424   Fabrika        :   56   Mean   :  187.7   Mean   :  40.94    Mean   : 15760
3rd Qu.:  475000   Ofis           :  897   3rd Qu.:  180.0   3rd Qu.:  33.00    3rd Qu.: 21340
Max.   :91500000   Residence      :  601   Max.   :22167.0   Max.   :1509.00    Max.   :100736
                   Villa          : 1196
district_population district_popularity
Min.   :   349      Min.   :    0
1st Qu.:452227      1st Qu.:    0
Median :828166      Median : 1459
Mean   :672971      Mean   : 4623
3rd Qu.:887915      3rd Qu.:10476
Max.   :887915      Max.   :10476
```

Summary of the target feature and predictor features