Intro to Data Mining Kaggle Presentation 2

YOUNAS MAHMOOD 22854

CRISP DM (1)

Business Understanding:

- This data set is of 'Property Price Prediction', we are solving this problem to understand how different features are important in guessing the price of properties
- By thorough research it will be very significant in understanding what increases/decreases the value of properties or is insignificant in increasing the value of properties.
- This will be beneficial for customers as through these predictions, they will be more confident in purchasing a property

Data Understanding

- Using Data.describe(), I understood the number of values each column has and whether it has any missing values, I also got better understanding of their mean, outliers, quartiles and standard deviation
- Using Data.corr(), I understood correlation of columns with each other, and I removed the highly correlated columns for better model execution time

CRISP DM (2)

```
data.describe()
```

✓ 27.1s

	floor	raion_popul	green_zone_part	indust_part	preschool_quota	preschool_education_centers_raion	school_quota	school_education_ce
count	98801.000000	100000.00000	100000.000000	100000.000000	86770.000000	100000.000000	87921.000000	
mean	6.968304	79841.65081	0.203707	0.152872	2487.148369	3.698700	6171.656521	
std	4.244951	45436.42718	0.179659	0.138926	1360.193047	2.318614	2706.690570	
min	0.000000	2546.00000	0.001879	0.000000	0.000000	0.000000	1012.000000	
25%	3.778876	41504.00000	0.063648	0.036122	1427.000000	2.000000	4474.000000	
50%	6.045461	80791.00000	0.137976	0.127376	2235.000000	4.000000	5824.000000	
75%	9.340616	106445.00000	0.317802	0.238617	3390.000000	5.000000	7653.000000	
max	33.348242	219609.00000	0.852923	0.521867	7610.000000	10.000000	16049.000000	

8 rows × 91 columns

```
data.corr()
```

✓ 7.4s

	floor	raion_popul	green_zone_part	indust_part	preschool_quota	preschool_education_centers_raion	school_quota	school_
floor	1.000000	0.011562	0.026846	-0.047508	0.082269	0.038932	0.076734	
raion_popul	0.011562	1.000000	-0.007551	-0.097418	0.796918	0.817484	0.871556	
green_zone_part	0.026846	-0.007551	1.000000	-0.497019	0.175514	0.022571	0.192562	
indust_part	-0.047508	-0.097418	-0.497019	1.000000	-0.278846	-0.103829	-0.301823	
preschool_quota	0.082269	0.796918	0.175514	-0.278846	1.000000	0.693489	0.856656	
...	...	...	...	...	...	...	...	
cafe_sum_5000_min_price_avg	0.027017	-0.048664	0.007929	-0.109002	-0.057334	-0.044913	0.064922	
cafe_sum_5000_max_price_avg	0.031606	-0.040569	0.005328	-0.116511	-0.040496	-0.037137	0.065520	
cafe_avg_price_5000	0.029858	-0.043755	0.006342	-0.113723	-0.047050	-0.040195	0.065340	
mosque_count_5000	-0.027850	0.046373	-0.219706	0.019418	-0.276592	-0.056418	-0.133417	
market_count_5000	-0.116292	0.322424	-0.302235	0.199918	-0.081591	0.268651	-0.063487	

91 rows × 91 columns

CRISP DM (3)

Data Preparation:-

1. For reading the 'Training set' and 'Test set', I used 'Numpy Excel Reader'.
2. I used numpy missing numbers method to identify the missing elements of columns, then I summed them and replaced them with the mean of the columns
3. I used numpy correlation method to identify which columns were having correlation above 0.95 or 0.9, then I removed those columns.
4. I used one hot encoding/ dummy encoding to convert the categorical columns to bits
5. I removed categorical columns having >3 or >2 attributes to reduce biasness
6. I tried to change outliers to median values of column through interquartile range
7. I also applied a variance threshold for some entries for better results

Modeling:-

1. I used a variety of models such as Random Forest Regression,Extra Tree Regression, Decision Tree Regression, Linear Regression, Gradient Boosted Regression, Stacking. Out of all these models, Extra Tree Regression was the best
2. Increasing the number of Estimators and criterions, increased the accuracy but slowed down processing time
3. Using different random states also improved the accuracy sometimes, but mostly I used the common random_state = 42
4. I tried to implement Hyper Parameter Tuning for Extra Tree Regression for better understanding of which parameters are most important for best score

Evaluation:-

1. Sometimes, I tested different conditions of models by splitting the training set, and checking the errors in python.
2. Mostly, I used the Kaggle scoring system to evaluate if the given entry was better than the last.

Entry	Pre-processing techniques	Model configurations	Kaggle Score	Understanding
1	One-hot encoding, removed categorical column with > 3 attributes. Removed rows with missing values.	Linear regression with normalization, default settings	7709036.78805	Did not correctly write output in Sample file
2	One-hot encoding, removed categorical column with > 3 attributes. Removed rows with missing values.	Linear regression with normalization, default settings	22463495.6163	Output in Sample file still not correct
3	One-hot encoding, removed categorical column with > 3 attributes. Removed rows with missing values.	Linear regression with normalization, default settings	22463495.6163	Previous sample file repeated
4	**One-hot encoding**, removed categorical column with > 3 attributes. Removed rows with missing values.	**Linear regression** with normalization, default settings	**5675903.49808**	Linear Regression gives a poor result
5	One hot encoding, **removed categorical column with > 3 attributes.** Removed rows with missing values.	**Gradient Boosted Regression**, default settings	**2542918.81343**	Gradient boosted regression, gives a fairly good result
6	One-hot encoding, removed categorical column with > 3 attributes. **Removed rows with missing values.**	**Random Forest Regression,** default settings	**1780574.76139**	Random Forest Regression, gives the best result yet
7	One-hot encoding, removed categorical column with > 3 attributes, **replaced missing values with column mean.**	**Decision Tree**, default settings	**2965753.21183**	Decision Tree regression, gives an average result
8	One-hot encoding, removed categorical column with > 3 attributes, replaced missing values with column mean.	Linear regression with normalization, default settings	3320100.06305	No normalization, and no removing of rows due to missing values, improved score.
9	One hot encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95	**Random Forest Regression,** default settings	1765850.34014	Correlation filtering improved score

Entry	Pre-processing techniques	Model configurations	Kaggle Score	Understanding
10	One hot encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95	Decision Tree, Default settings	2957875.29684	Correlation filtering improved score of Decsion tree as well
11	One hot encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95. Removed rows having outliers	Linear regression with normalization	3513515.01412	Removing rows of outliers decreased score.
12	One hot encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95.	**AdaBoost Regresion**	3075196.04853	Average score by Adaboost
13	One hot encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95.	Decision Tree, Splitter changed to Random	2994590.38819	Splitter best was better instead of random
14	One hot encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95. Removed rows having outliers	Gradient Boosted Regression	2880326.99357	Removing outliers was of no use
15	One hot encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95.	Extra Trees Regression	1580327.38598	Extra Tree Reression was best model yet
16	**Dummy encoding**, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95.	Extra Trees Regression with N_Estimators = 1200 RandomState = 42	**1576356.07399 (PB)**	N_estimators = 1200, Dummy Encoding, improved score
17	Dummy Encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95. **Forward selection** with n_features = 30.	AdaBoost Regression	3219103.834	Forward Selection didn't improve score

Entry	Pre-processing technniques	Model configurations	Kaggle Score	Understanding
18	Dummy Encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95. **Forward selection** with n_features = 30.	Random Forest Regression, n_estimators = 400	1916517.91992	Forward Selection doesn't improve score for random forest
19	Dummy Encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95. **Forward selection** with n_features = 30.	Extra Tree Regression, n_estimators = 400	1651756.60957	Forward Selection doesn't improve score for Extra Tree
20	Dummy Encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95. **Forward selection** with n_features = 30	Decision Tree Regression	3223007.65802	Forward Selection doesn't improve score for Decision Tree
21	Dummy Encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95. **Forward selection** with n_features = 30	Linear Regression	3228566.32706	Forward selection doesn't improve score for Linear Regression
22			7709036.78805	Error in output
23	Dummy Encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95. **Forward selection** with n_features = 30. Filled Outliers with column median values	Linear Regression	3639124.53982	Not the best approach to deal with the outliers in data
24	Dummy Encoding, removed categorical column with > 3 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95. Applied a **variance threshold** of 0.99	Random Forest Regression, n_estimators = 400	1776011.68718	Variance threshold improved score marginally

Entry	Pre-processing techniques	Model configurations	Kaggle Score	Understanding
25	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95.	Extra Tree Regression n_estimators = 500, random_state = 42	1587800.55389	Removing column 'Ecology' had about the same impact.
26	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95.	Extra Tree Reg n_estimators = 500. random_state = 42 criterion ='mae'	1692419.90303	Criterion changed to mae from 'mse'
27	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95.	Voting regrerssor of Extra Tree Reg and Random Forest Reg	1778145.45141	Stacking did not improve score a lot.
28	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95. **Backward Selection** n_features = 50	Linear Regression	3224009.72252	Backward selection was useless because it took too much time for the same result
29	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.95. Backward Selection n_features = 50	Extra Tree Reg n_estimators = 50. random_state = 42	1602789.2304	Same type of result for backward selection as forward selection
30	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.90.	Extra Tree Reg n_estimators = 250. random_state = 42, Depth = 10	1586881.3144	Same type of scores as before for Extra Tree Reg

Entry	Pre-processing techniques	Model configurations	Kaggle Score	Understanding
31	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.90.	Voting regressor of Decision Tree and Linear Regression	2959412.5331	Confirming voting regressor of Decision Tree and Linear Regression
32	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.90.	Extra Tree Reg n_estimators = 1400.	1583855.76679	Same type of score with little increase in n_estimators
33	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.90.	Extra Tree Reg n_estimators = 2500	1583109.82449	Increasing n_estimators much more, still did not improve score much
34	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column median. Removed columns have correlation > 0.90.	Decision Tree Regressor	3141048.77967	Replacing missing values with mean is better than median
35	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.90.	Random Forest Regressor n_estimator = 400	1769430.25773	Confirming random forest individually after removing column 'Ecology'
36	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.90.	Linear model with forward selection of n = 40	3224011.18253	Poor result using forward selection
37	Dummy Encoding, removed categorical column with > 2 attributes, replaced missing values with column mean. Removed columns have correlation > 0.90	Random Forest Regression with forward selection of n = 40	2034422.16755	Average result due to forward selection

Overall findings & insights

- The best model for the dataset for me was definitely 'Extra Tree Regression’, because unlike 'Random Forest' it draws samples without replacement which reduces repetitions of observations. Also, it's splitting was done randomly which reduces variance.
- Label Encoding was the most fruitful method in Kaggle Score among other Encodings
- Removing columns through correlation, helped in execution time, and also improved the score of models
- The challenges I faced were finding the best model for the score, which I found out through trial and error. After finding the best model, I then started finding the best configurations of that model to improve the score through 'Hyper Parameter Tuning'.
- After research about the best pre-processing techniques, I found out about KNN Imputation for missing values, and to replace outliers with medians of columns.
- The biggest challenge was trying to execute backwards selection, polynomial interactions and polynomial regression, out of which only backward selection with less columns could be completed in my PC.