Name:- Swapna Vaidya ***Class:-*** MSDS – 422

COLAB NOTEBOOK LINK :- https://colab.research.google.com/drive/1YJGfK6neilEktn98OeU3nK3zCzK1CCjQ

Evaluating Classification Models

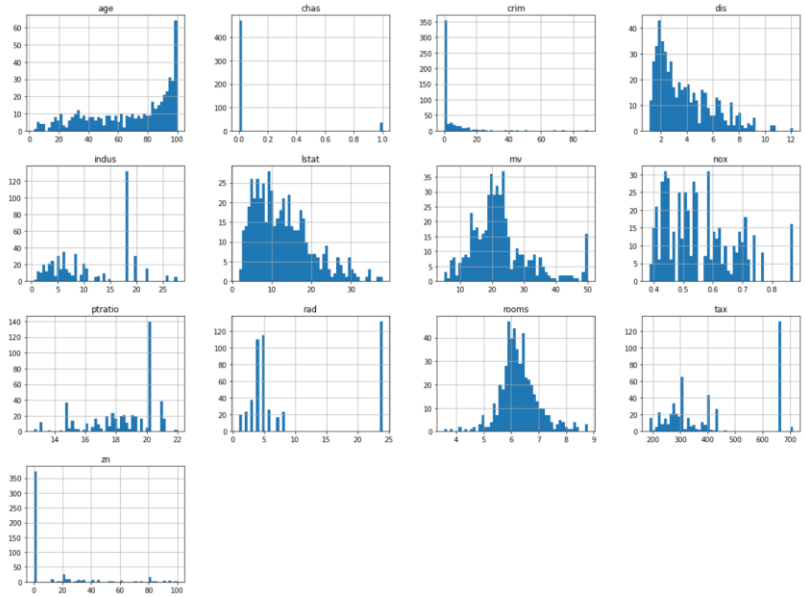
(1) ***A summary and problem definition for management***

The Boston Housing Study is a market response study of sorts, with the market being 506 census tracts in the Boston metropolitan area. The objective of the study was to examine the effect of air pollution on housing prices, controlling for the effects of other explanatory variables. The response variable is the median price of homes in the census track. The real estate firm wants to use machine learning to complement conventional methods for assessing the market value of residential real estate. Of the modeling methods examined a recommendation needs to be made to the management on the model and the reason for it.

(2) ***Discussion of the research design, measurement and statistical methods, traditional and machine learning methods employed***

Based on the exploratory data analysis that was done the survey data, below were some of the findings:-

- In all there are 506 instances with 13 features in the data
- Of the 13 features, the neighborhood feature will not be used for the analysis.
- All the 13 features are non object features.
- There are no missing values in any of the columns in the data set.
- After looking at the distribution of the features, none of the features except “rooms” has a normal distribution. The feature “age” is left skewed with some extreme oultiers between the age of 90 and 100. Similarly for the feature “chas” majority of the residents do not live on the Charles river. The feature “crim” shows that right skewed distribution with majority crim rate being zero and overall crime rate being low. The feature “dis” shows that the distribution is right skewed with majority of the residents living near the employment center. Similarly “tax” distribution is also right skewed with some big outliers for the higher tax rate. Finally the “zn” feature distribution tells that majority % of the land is not zoned for lots.



- Based on the correlation matrix of the 12 features, we can see that the feature “dis” which indicates the weighted distance from the employment centers has the lowest correlation with the median house price. Similarly the feature “chas” whether the house is on the river Charles or not has the lowest positive correlation with the median price of the house. While the feature “rooms” which indicates the number of rooms in the house has the highest correlation with the median price of the house. While the feature “lstat” which indicates the percentage of population of lower socio-economic status has the highest negative correlation with the median price of the house. The lower the % of population of lower socio-economic status the higher the median price of the house. Rest of the feature either have positive or a negative correlation with the response variable.

	crim	zn	indus	chas	nox	rooms	age	dis	rad	tax	ptratio	lstat	mv
crim	1.000000	-0.200469	0.406583	-0.055892	0.420972	-0.219247	0.352734	-0.379670	0.625505	0.582764	0.289946	0.455621	-0.389582
zn	-0.200469	1.000000	-0.533828	-0.042697	-0.516604	0.311991	-0.569537	0.664408	-0.311948	-0.314563	-0.391679	-0.412995	0.360386
indus	0.406583	-0.533828	1.000000	0.062938	0.763651	-0.391676	0.644779	-0.708027	0.595129	0.720760	0.383248	0.603800	-0.484754
chas	-0.055892	-0.042697	0.062938	1.000000	0.091203	0.091251	0.086518	-0.099176	-0.007368	-0.035587	-0.121515	-0.053929	0.175663
nox	0.420972	-0.516604	0.763651	0.091203	1.000000	-0.302188	0.731470	-0.769230	0.611441	0.668023	0.188933	0.590879	-0.429300
rooms	-0.219247	0.311991	-0.391676	0.091251	-0.302188	1.000000	-0.240265	0.205246	-0.209847	-0.292048	-0.355501	-0.613808	0.696304
age	0.352734	-0.569537	0.644779	0.086518	0.731470	-0.240265	1.000000	-0.747881	0.456022	0.506456	0.261515	0.602339	-0.377999
dis	-0.379670	0.664408	-0.708027	-0.099176	-0.769230	0.205246	-0.747881	1.000000	-0.494588	-0.534432	-0.232471	-0.496996	0.249315
rad	0.625505	-0.311948	0.595129	-0.007368	0.611441	-0.209847	0.456022	-0.494588	1.000000	0.910228	0.464741	0.488676	-0.384766
tax	0.582764	-0.314563	0.720760	-0.035587	0.668023	-0.292048	0.506456	-0.534432	0.910228	1.000000	0.460853	0.543993	-0.471979
ptratio	0.289946	-0.391679	0.383248	-0.121515	0.188933	-0.355501	0.261515	-0.232471	0.464741	0.460853	1.000000	0.374044	-0.505655
lstat	0.455621	-0.412995	0.603800	-0.053929	0.590879	-0.613808	0.602339	-0.496996	0.488676	0.543993	0.374044	1.000000	-0.740836
mv	-0.389582	0.360386	-0.484754	0.175663	-0.429300	0.696304	-0.377999	0.249315	-0.384766	-0.471979	-0.505655	-0.740836	1.000000

(3) ***Overview of programming work***

This assignment was done in Google Colab, using 4 main libraries (Pandas, numpy, matpotlib, Seaborn and Skylearn).

- ***Ingestion:-*** Using pandas the data was first loaded from Google OnDrive. The feature “neighborhood” was removed from the dataset
- ***Exploration:-*** of the data was done using methods like describe(), info() and corr().
- As part of exploration, ***visualization*** was done using histograms and scatterplots.
- ***Data Cleaning and Transformation:-*** No special data cleaning and transformation steps were taken on this data set.
- ***Modeling*:-** The label “mv” was separated from the feature value data set. The feature data set was then standardized using the “StandardScalar” method. Using a 3 fold cross validation methodology 3 different model were used for training the data set. The first model used was the “Ridge Regression” with solver = “Cholesky” and the alpha parameter set to 10. This resulted in a RMSE of 6.70 and an adjusted Rsquare of 46.70%. The same model with the solver = sag and alpha parameter set to 100 was also used. This resulted in a RMSE of 5.90 and an adjusted Rsquare of 58.55%. Next a “Lasso” model was used with an alpha of 0.1. This resulted in a RMSE of 5.90 and an adjusted Rsquare of 48.05%. Finally the “Elastic Net” model was used with an alpha of 0.3 and the ratio of 0.5. This resulted in a RMSE of 5.90 and an adjusted Rsquare of 56.84%.

(4) ***Review of results with recommendations for management***

After the exploratory data analysis and modeling exercise has been completed, following are the recommendations to the management of the real estate agency:-

- Of the 3 models used, the Ridge Regression Model with the solver = sag with alpha set to 100 should be used. This gives the lowest RMSE of 5.90 with a highest R square of 58.55%. This will ensure that bias is low and the accuracy of the median price predication for new houses will be high. Since we are using the Ridge Regression model, with alpha set to 100 the insignificant features will be almost zero but not zero hence the model will still be complex resulting in some variability.

  Using this model a sample of the predictions compared to the actuals are below. The data could be further cleaned and transformed to take care of some of the extreme outliers as we can see that when the price is the highest the mean square error for those instances is high.

	Org House Price	Pred House Price	difference
0	24.0	32	-8.0
1	21.6	25	-3.4
2	34.7	29	5.7
3	33.4	29	4.4
4	36.2	27	9.2
5	28.7	27	1.7
6	22.9	24	-1.1
7	22.1	18	4.1
8	16.5	9	7.5
9	18.9	19	-0.1