Topic 1: Machine Learning Packages in R And Python

Jonathan Ibifubara Pollyn

GitHub Repository: https://github.com/JonathanPollyn/Machine-Learning-for-Data-Science

College of Science, Engineering and Technology, Grand Canyon University

DSC-540: Machine Learning for Data Science

Dr. Aiman Darwiche

November 03, 2021

Prediction of Real Estate Appreciation Over Time

The prediction of real estate appreciation over time was performed using the housing dataset with a total of 20640 observations, as shown in figure 1, but there are missing records from the total bedroom. From the data overview in figure 2 shows that the datasets attributes are all numerical values except the ocean proximity.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

Figure 1: Data information

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

Figure 2: Housing data set

Figure 3 shows that the maximum median house value is 500001, which indicates the maximum house value at the time of this data, with the maximum housing median age of 52 years.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000	20640.000000
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816909
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615874
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.000000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000000
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.000000
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000

Figure 3: Description of the house data

The housing dataset is split into a training and test data set. The test data set is used to validate the model. Multiple regression analysis was performed on the dataset by making the housing median age, population, and total rooms the predictor variables and the median house value as the response. This is to analyze how age, population, and total rooms affect the house price values. Figure 4 shows that all the variables used in the model are statistically significant because they are all below the p-value of 0.05. The model's coefficient is 143607.21 where the median housing age, population, and total rooms have a coefficient of 1704.18, -55.56, and 35.61 subsequently. The coefficient for the housing median age indicates that while keeping everything constant, a 1 unit increase in the housing median age will be associated with an increase of 1704.18 dollars in median house value. The population coefficient states that while keeping everything constant, a 1 unit decrease in the population will decrease the median house value by -55.56; keeping everything constant 1 unit increase in total rooms will increase the median house value by 35.61 dollars.

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	143607.2145	2795.4275	51.3722	0.0000	138127.8756	149086.5534
housing_median_age	1704.1759	71.7420	23.7542	0.0000	1563.5538	1844.7980
population	-55.5579	1.4434	-38.4912	0.0000	-58.3871	-52.7287
total_rooms	35.6111	0.7739	46.0143	0.0000	34.0941	37.1280

Figure 4: Summary of the model

Model validation

The model was validated using the test data. Figure 5 shows that the housing median age, population, and total rooms have a coefficient of 1696.06, -50.83, and 32.35 subsequently. The median housing age and total rooms show approximately the same increase, and the population shows a decrease of 50.83 dollars in median house value. The model's mean absolute error (MAE) for the regression values was calculated to be 84442.49. The mean absolute error (MAE) baseline was also calculated to be 83987.30 because the mean absolute error (MAE) regression values are greater than the MAE baseline value indicated that the model's mean absolute error (MAE) for the regression outperforms the baseline mean absolute error.

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	144482.5362	5622.6270	25.6966	0.0000	133459.1545	155505.9180
housing_median_age	1696.0575	143.7947	11.7950	0.0000	1414.1423	1977.9727
population	-50.8251	2.8976	-17.5405	0.0000	-56.5059	-45.1443
total_rooms	32.3592	1.4892	21.7292	0.0000	29.4396	35.2789

Figure 5: Model Summary of the test data

Conclusion

The prediction explains that real estate properties will appreciate more in districts where the properties have more rooms. The prediction also shows that as districts grow in age, real estate value will appreciate, however as the population declines in some districts, real estate values will depreciation.

Using Machine Learning to Translate Applicant Work History into Predictors of Performance and Turnover

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, *104*(10), 1207–1225. https://doi.org/10.1037/apl0000405

According to the authors, to screen candidates, employers frequently use resumes and job application forms that include information about previous employment. The authors affirmed that there is little agreement on how to systematically translate past employment information into indicators of future employment outcomes. The authors stated that machine learning techniques are used to generate easy-to-read measures of work experience relevance, tenure history, history of involuntary turnover, and a history of staying away from bad jobs and approaching better positions, using data from job application forms. When applied to a longitudinal sample of 16,071 public school teaching positions, the authors stated that the model accurately predicts future work outcomes such as student evaluations and expert performance observations and value-added to student test scores. Having relevant work experience and a history of approaching better positions were found to be associated with beneficial work results, whereas staying away from less desirable occupations was linked to less favorable results." Furthermore, the authors stipulated that estimating the amount to which the approach can increase the quality of the selection process compared to current techniques of assessing job history while reducing the likelihood of negative consequences.

The approach by the authors classified self-reported job titles and descriptions into a standardized occupation code using supervised machine learning techniques. Then when training an algorithm on a large, external dataset, the authors affirm that it's best to use supervised classification to ensure the results are as accurate as possible. Supervised learning appears to be a good approach considering Hastie et al. (2001) stipulated that the goal of supervised learning is to learn from a teacher's example. A training set of observations is compiled by observing the system under investigation, including inputs and outputs. This artificial system, called a learning algorithm, uses the observed input values to produce outputs in response to the input values. Due to input/output relationship changes, the learning algorithm can adjust its learning curve to account for these.

The authors also used the naïve Bayes classifier to train the occupational descriptions and job titles algorithm, using full job descriptions and alternative job names for 974 different jobs to train the classifier.

Conclusion

As much as the method produced the needed result in this situation, there could be an issue using this method. Obtaining data of applicants from their previous work experience could be difficult as most applicants can fill out a form providing the information, they believe you need to know. Above all, HR information is not always straight forward as HR data could be bias. The authors affirmed that training a large data set in a supervised classification will produce the best result. It is important to know that, as stated by Hastie et al. (2001), the goal of supervised learning is to learn from a teacher's example. A training set of observations is compiled by observing the system under investigation, including inputs and outputs. This artificial system, called a learning algorithm, uses the observed input values to produce outputs in response to the input values. Due to input/output relationship changes, the learning algorithm can adjust its learning curve to account for these. If we do not have good quality data to feed the model, we could make the wrong decision. According to Hastie et al. (2001), the Bayes Classifier is particularly useful when the feature space's dimension p is large, as this makes density estimation undesirable. According to the naïve Bayes model, if G = j, then features Xk must be independent.

References List

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer New York. https://doi.org/10.1007/978-0-387-21606-5

Larose, C. D., Larose, D. T., & Larose, Chantal D., Author. (2019). *Data science using python and r*. John Wiley & Sons,inc,.

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology, 104*(10), 1207–1225. https://doi.org/10.1037/apl0000405