Multiple linear regression

Problem

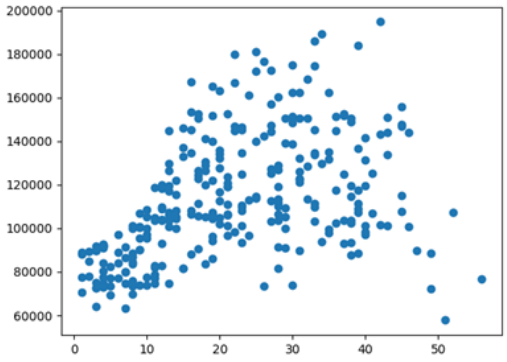
Goal of this program is to find a regression model that best fits the data set, which contains X and Y values. X values are title, field, years as a doctor, years of experience, and gender. Y value represents yearly salary.

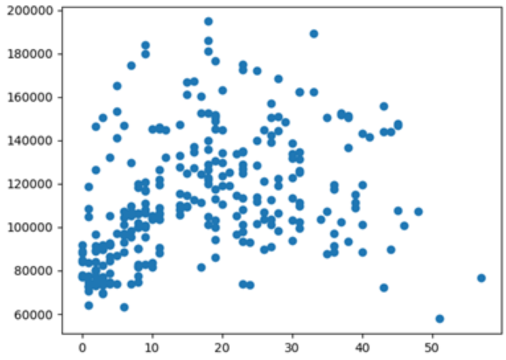
Data

Since in our data set we have multiple variables that affect salary, we decided to inspect the data set by each variable to see if we find any regularity. We calculated the average salary by each category.

Assistant Professor	Associate Professor	Professor
80 591	95328	126820

As expected professors have the highest average salary while assistant professors have the lowest. This helps us better interpret the results we later got using ridge regression. Even though years as a doctor and years of experience are obviously highly correlated we decided to keep both variables to check how our models interpret them. Visualizing these two correlated to salary we found no obvious correlation therefore we can conclude other marks have more impact on salary.





We can notice that the group with less than 10 years as a doctor has a lower salary overall but after that, no real rule can be deducted. Also, it is worth mentioning that the title mark is correlated to both of these marks(most of the assistant professors have less than 10 years as a doctor) and more about this will be explained in the outlier section.

Ridge regression

We implemented a closed form solution of the ridge regression algorithm. We used multiple values for the l2 penalty and by cross-validating (k=10) we got the best result with l2 = 1.5.

On normalized marks (using z-score) our theta result looked like this:

Intercept	Gender	Field	Title	Years as a doctor	Years of experience
314.48296613	8.92200765	20.54049939	28.00396217	5.88110532	-5.22794801

From this we can deduce that the highest impact on salary has a title mark and that years of experience and years as a doctor have a very strong correlation.

KNN

Another approach we tried was KNN. K that provided the best results in cross-validation was 20. The distance metric that we used is Euclidean distance.

Target encoding

Target encoding is a Baysian encoding technique.

Bayesian encoders use information from dependent/target variables to encode the categorical data. In target encoding, we calculate the mean of the target variable for each category and replace the category variable with the mean value.

In the case of the categorical target variables, the posterior probability of the target replaces each category.

Outliers

By plotting data we can see that there is only one professor with no years of experience so we decided to remove that observation. Next, we focused on removing professors with very high salaries since our model usually gave higher predictions. After removing all professors that have a salary greater than 190000 and more than 35 years of experience, our model gave all around better predictions (equally higher and lower). For associate and assistant professors, we removed those who have a salary lower than 72000 (associate professor) and lower than 73000 with no years of experience (assistant professor). Overall we removed 6 observations(0.02% contamination).

Results

Ridge, k fold cross-validation (k = 10)

	One hot encoding	Label Encoding	Target Encoding
Average RMSE	20413	20546	20394

KNN, k fold cross-validation (k = 10)

	One hot encoding	Label Encoding	Target Encoding
Average RMSE	22635	22786	20252

Besides this we also tried z-score and min-max normalization but got no improvement. We tried multiple normalizations on the target variable in order to lower skewness but also without achieving better results.

Conclusion

Using KNN and Target encoding we achieved RMSE of 22797.77 on the final test set.

References

https://medium.com/where-quant-meets-data-science/building-k-nearest-neighbour-algorithm-from-scratch-bd0c5df13192

https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

https://towardsdatascience.com/dealing-with-categorical-variables-by-using-target-encoder-a0f1733a4c69

https://machinelearningmastery.com/model-based-outlier-detection-and-removal-in-python/

Filip Volarić SW54-2018

Svetozar Vulin SW57-2018