21-12-2020

Machine learning Group Project

By Seung Woo Hong (10879420), Jinhyun Kim (11968850), Ee June Kim (11630566), Noemi Roos (12410292)

University of Amsterdam

Minor: Amsterdam Data Science and Artificial Intelligence

Machine Learning

Data

With the prevalence of depression becoming more and more common, it comes to no surprise that researchers have been spending more time trying to understand why this phenomenon is increasing and how it even comes about in the first place. There could be many reasons for the rise, whether due to the increased use of social media, genetic factors or environmental reasons.

This report concerns itself with the CBS data set that encompasses statistical data, primarily demographic, social and geographical trends for all neighbourhoods in the Netherlands. This data set was combined with the RIVM depression data set to further set the foundational groundwork in order to explore, analyse and extrapolate trends and to ultimately say something about how depression comes about. The analysis aims to shed insight into the multifaceted ways depression materialises and tries to pinpoint the likely causes of it using statistical models mainly revolving around regression analysis.

Preprocessing

As previously mentioned, the CBS as well as the RIVM data was combined, so that both the target variable ("Matig of hoog risico op angst of depressie (%)") and all the feature variables are in a single data set. This was done by merging the data on the "Gebied" variable. Additionally, certain features pertaining to area codes or regions were dropped since categorical features will not be used in the selected models. Whilst inspecting the data, some inconsistencies were observed. Firstly it was noticeable that there were many missing values, many of which taking the form of ".". These missing values were dealt with in several ways. When there were more than 100 missing values in one column, this column was deleted since without more than 100 missing cases, it was hard to use the column in the model. While the missing values were under 100, the mean of the column was used to deal with the missing values. Additionally, all values were converted to floats for more consistent computing. In terms of the decimal measurements, the merged dataset had both decimal points as well as decimal commas previously belonging to the depression and CBS dataset respectively, so all decimal commas were converted to decimal points. In order to properly conduct our additional analyses for the second half of the project seen in the coming sections, columns were renamed and the necessary computations were performed to obtain all values as ratios. Due to the relationship between variables, scaling cannot be used. It is because each ratio could be misinterpreted and lose the relationship between the values while using the scaler. Therefore, the random forest, which does not affect the relationship between the ratio, was additionally used as regression.

EDA

In order to explore our data, each of the variables were plotted on a histogram to showcase possible irregularities. Results of the feature distributions demonstrated that a majority of the feature variable were not symmetrical but were primarily skewed to the left. This makes sense as the majority of values are closer to zero. Examples of less skewed data include the “p_koopw”, “p_huurw”, “g_hhgro variables”. In terms of the target variable, it seems to be fairly symmetrical around 40. One may assume that scaling may prove to be beneficial.

Models

Using original data

For the following models, we firstly use the original data set that contained all values in the form of counts. Analysis-wise, three regression models were settled on that were deemed most suitable for this data analysis task, namely linear regression (OLS), ridge as well as lasso.

1. *Ordinary least squares*

   OLS does not have any parameters to tune so as a result does not allow for complexity to be controlled for. This can lead to overfitting in data sets with large numbers of features as is the case with this dataset. However, it may prove useful to initially use a simple regressor to explore the data and get acquainted with feature coefficients. General results gained from this model are training and test set scores being fairly close at 0.409 and 0.391 respectively, which may point towards underfitting. Without tuning the regressor and simply using 10 fold cross validation, we obtain an R squared of 0.384 with an MSE of 26.1 and an MAE of 4.1. After using a pipeline to introduce different scaling options, namely MinMaxScaler, StandardScaler or passthrough (no scaling), a score of 0.41 was obtained with best scaler being MinMaxScaler - a slight improvement to the previous score is observed.

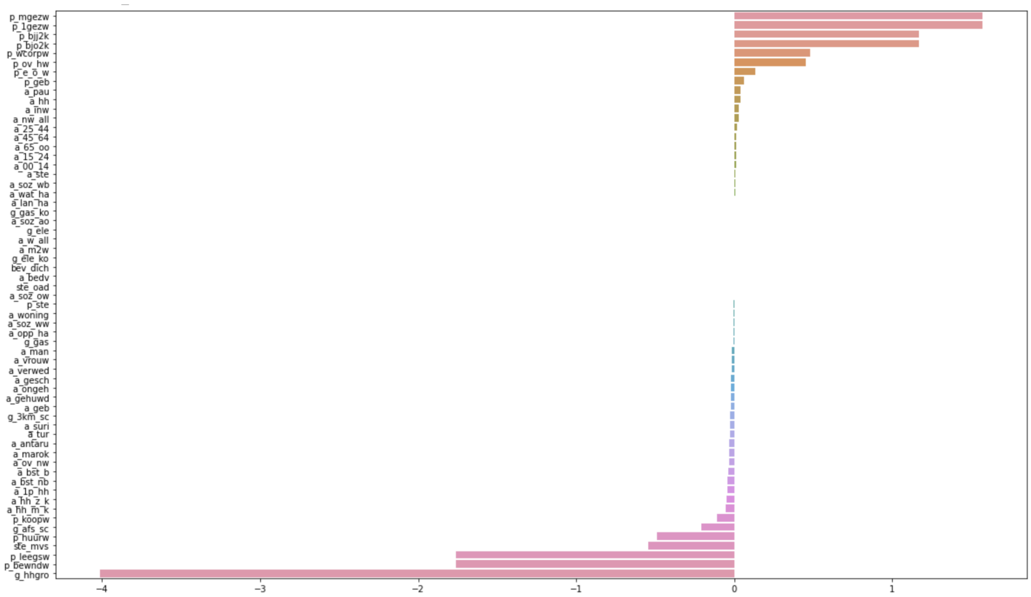

Figure. 1 coefficient magnitudes of linear regression

The above figure shows us the features that have the most effect on our target variable. Here, the highest coefficients come from p_mgezw and p_1gezw. The most strongly negative one comes from g_hhgro with a magnitude reaching -4. It is also obvious that a majority of the features have a fairly minimal effect.

2. *Ridge*

By turning to ridge, features can be controlled for using L2 regularization with near zero coefficients by using the hyperparameter alpha. With this in mind, we hope to further decrease overfitting in hopes of improving results. Very similarly to OLS, here a non-tuned regressor gives a training score of 0.41 and a test score of 0.39 and 0.38 using 10 fold cross validation. Using a pipeline, the best scaler was StandardScaler with a score 0.41 and with the best alpha parameter at 0.21. Even with tuning, results seem to be similar. After adding polynomials and increasing feature size from 62 to 2015 variables and using the best parameter at alpha = 1000 with no scaling, the best training and test score was obtained at 0.48 and 0.39 with an MSE of 23.2 and an MAE of 3.8. Generally, ridge seemed to perform slightly better than OLS. Since alpha was set high at 1000, it can be assumed that many coefficients were pushed towards zero. Additionally, since training and test scores are further apart, one could assume that the model is overfitting with polynomials included.

3. *Lasso*

By using L1 regularisation, lasso may also be used as a preprocessing step by determining only the most crucial features. Results from lasso give a training score of 0.32 and a test score of 0.33 with polynomials. Since lasso conducts feature selection, the model uses 16 of the

2015 features. However, these scores are even lower than that of ridge, so polynomials were excluded giving an increased training and test score of 0.39 and 0.39 with 45 of the original 62 features used. MSE and MAE give a score of 27.3 and 4.2 respectively. Using a pipeline, ideal alpha at 0.01 and ideal scaling with “passthrough” was obtained along with a best score of 0.40 - the best score achieved with lasso.

Ultimately, results did not seem to deviate too much with all three of the models. The best performing model was ridge with an R squared score of 0.48 when setting alpha = 1000 and including polynomial features. However, all models still give a small R squared, which indicates that all the models used so far are still weak.

Using recomputed data to ratios

_______Since scaling did not have much of an effect in the previous models, an alternate approach was taken. As previously mentioned in the preprocessing step, for the recomputed data, all values were recalculated to allow for more consistent outcomes since values in form of counts might skew results in an undesirable way.

1. *Random forest regression*

   The random forest is one of the learning methods for classification. The random forest is a well known regressor that has high accuracy, fast test algorithm, and can avoid overfitting of the model. First, We used test size 0.2, and random state 100 for the base model. The root mean square deviation (RMSE) of this base model is 4.97. Second, to delete the outliers we used an isolation forest. After deleting the outliers the RMSE socore became 4.96. Next, we did the feature selection to reduce the complexity of the model and to make the interpretation of the model easier. In this step, we decide to use only 40 best features in the model which can help the model to find out the relationship between the depression and the variables. However, there was no improvement in feature selection, we used whole features with out outlier deleted. The hyper parameter tuning was done to find out the best parameters which were max depth as 7, max feature as sqrt, min sample split score as 0.5 and N estimators score as 500. Lastly, the random forest regressor for this model shows the increased RMSE score by 5.61

2. *XG Boost Regressor*

   We also used XG boost regressor, one of the Machine Learning algorithms that is based on a decision tree. This method is usually used to solve the classification, ranking, regression and user defined prediction problems. Similar to the random forest model, we used test size 0.2, and random state 100 for the XG boost model. In the first model, we got 4.95 as the RMSE score. Again, the deleting outliers had been done with isolation forest. With the outlier deleting, the RMSE score became 4.92. The feature selection which we used 40 best features for the model showed RMSE score 4.92. This feature selection had no influence on the model, we did the hyper parameter tuning by gridsearchcv. From this hyper parameter tuning we found the best parameters which were colsample by tree as 1, learning rate as 0.05, N estimators as 500 and subsample as 0.6. The final model with best parameter results 4.89.

Lastly, we did the neural network model to check the models. With this model, the overfitting was not observed. Also, it was found that there was less data to perform better than tree based models.

Discussion

Since mean imputation was used, variability in the data may be reduced and thus the standard deviation may be minimised. Additionally, if the data contains a significant proportion of outliers, this may skew our data and make it less representative of the true values when means are imputed for missing data. Also, since the data had lots of missing values, it was difficult to use all the data. Therefore, for the future machine learning, it would be important to have less missing values. Another important point in this research is that due to the relationship between the values, the scaling was difficult to be used in the model. There were risks that if we use the scalar the meaning of the ratios would be misinterpreted. Therefore, in the research, we did not use any scalar.

Conclusion

With the CBS data, the research has been conducted to find the regression model that shows how the depression rate differs due to the population characteristics such as status of marriage, and racial. Since the data had limitations which are mentioned above in the discussion section, this research is not a perfect model. However, with the data we have conducted 5 models which could be divided into two parts. First part is using the original data which had the OLS, Lasso and Ridge model. These all three models showed low R squared scores. The highest R squared score among three models was ridge model scored by 0.48. Since the first part of models were not performing well, we made the second part of the models which were random forest model and XG boost model. These models used the recomputed data. For both models, we did outlier deleting with isolation forest and feature selection to find the best performing model. However, the feature selection of both models did not make any improvement on the model. Therefore, we did the hyper parameter tuning. After we found out the best parameters for the model, the RMSE of random forest model ended by 5.61 and the

RMS of the XG boost model scored 4.89. To conclude, the best performing model among the models that we made was the random forest model. With the random forest model, the depression rate change due to the population characteristic could be tracked.