COMP47350 Homework 2

Individual Report

Cian Belton 19321726

1. Personal contribution (500 words)

Me and Shuya decided to work on Exercise 0 and 1 together so that we both had to understand the problem in hand before moving to the next exercises. I wrote all of the code and markdown for Exercise 0, except the drop_feature and drop_column functions. I inserted the code from Homework 1 and integrated it with our merged dataset. This gave us a cleaned and merged dataset which was ready to be split into training and testing in Exercise 1. We used the sample solution provided as inspiration for this homework throughout.

The reason we went with my method of data preparation and cleaning was that my Homework 1 grade was an A, so we decided to follow my cleaning process for the data and add in the features that I created.

I wrote the code for plotting the bar charts in Exercise 1. I also interpreted these results in the markdown cells. This gave me a better understanding of the impacts of some features on death_yn. I wrote the code to see the correlations between the different features and the target feature. I fixed an error in the label encoding, by changing “North” to “Other” as Alaska and Hawaii fell into this category. It is true that Alaska is in the North, but Hawaii is not.

I have a good eye for detail, consistency and structure so I laid out Exercise 0 and 1. As I was still formatting and coding Exercise 1, Shuya moved on to exercise 2, 3 and 4. He wrote most of the code for these parts. However I revised and restructured a lot of these sections to flesh out our analysis, standardise our layout and make each exercise have a consistent feel. I fixed and checked spelling and grammatical errors. I revised the discussion of results for the confusion matrix to better explain what information was being conveyed by the matrix. I made the confusion matrix run on all training data not just the first 10 rows.

Once I had finished revising Exercises 2-4, I could finally move on to the final part of this homework. In Exercise 5 my goal was to improve the predictive models that we had created in the previous sections. I wrote all the code and markdown in Exercise 5. I found it challenging but rewarding to compare the random forest to the other classification models as I had to research which ones to include and then how I would compare them.

I researched the problem domain by reading a study about how age has such a large impact on deaths related to Covid-19. [1]. We used GitHub for software development so that we could work on the same notebook simultaneously.

My personal contributions helped the project to be successful as it laid the groundwork for data understanding and preparation and the analysis of the models Shuya wrote. I also wrote the readme.txt file that layed out our file structure for this homework.

2. What did you learn from the project? (500 words)

In Exercise 1 I saw how the target outcome was related to the different features in our dataset. In order to do this I learned how to code the bar charts and train test split.

In Exercise 2 I broadened my understanding of linear regression as I had to evaluate and improve upon what Shuya wrote. I learned more about confusion matrices and their benefit in plotting the 4 combinations of results for an outcome. I had to learn about the different measures used to evaluate the confusion matrix such as:

1. Accuracy: % correct predictions out of the total number of predictions.
2. Precision: % correctly predicted positives out of the total number of predicted positives.
3. Recall: % correctly predicted positives out of the total number of actual positives.
4. F1 score was just the mean of precision and recall.

I learned about cross validation and how it is used to obtain an unbiased estimate of the model's performance on new, unseen data. It also plays a vital role in reducing skewed evaluation metrics. It does this by running the regression multiple times and then getting the mean of the results.

In Exercise 3, I researched logistic regression and saw how much more effective it was at handling outliers than linear regression. It came as no surprise to me when this model performed better than linear regression at predicting the target outcome.

In Exercise 4, I learned about how a random forest works. I found it interesting that it uses majority voting to decide on the most suitable class as there is parallels to majority voting everywhere in society (elections, referendums and shareholder meetings). I expected this model to perform the best, and it did. This is because it uses a more intensive form of generating predictions than the previous two models.

In Exercise 5 I researched the different ways to improve classification models. Some of the methods were: review/include existing features to get the best combination, add new features, improve the random forest model by changing the maximum depth and number of estimators. The other sklearn models I learned about for comparing to the random forest model were:

1. Ada Boost: A boosting algorithm that combines multiple weak classifiers to improve the overall classification performance.
2. K Neighbours: A non-parametric algorithm that classifies new data points based on the k-nearest neighbours in the training data.
3. MLP: A neural network algorithm that learns to classify data by adjusting the weights and biases of multiple layers of interconnected neurons.
4. Gradient Boosting: A boosting algorithm that iteratively trains decision trees to correct the errors of the previous iterations, resulting in a strong ensemble model.

We worked very well on this project as we split up the parts and worked on what suited us best. Shuya writing the code for the models and laying the groundwork, and me understanding, evaluating and writing about them.

Time could have been saved in formatting the notebook if we had been more clear on what format to stick to at the start. Shuya did improve on this after some feedback from me which saved me having to reformat Exercise 4. I would avoid these mistakes in future by being more clear with structure and layout.

3. Anything else related to this project that was not covered in the previous 2 parts? (500 words)

I really enjoyed this project. It was great to put the concepts I learned about during the semester in practice on a massive amount of data. I was proud of how well our model can predict the target outcome with 11 of the original features and also 6 that I deemed to be of high and medium importance.

I really enjoyed writing about the results of the different predictive models on our dataset . I think it is amazing that we can get 94% predictive accuracy on death_yn using our notebook!

I found it time consuming but enjoyable to format and be meticulous working down through the different exercises in the notebook. I was very surprised by random forest performing better than all of the other classification models that I ran it against. I learned a lot by undertaking this homework and it has piqued my interest in considering Data Science as a career for when I graduate!

4. Further Reading I conducted:

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7247470/