**Capstone (DATA 2206)**

**Module Three - Preliminary Report**

**CancerPScreenRaw.csv**

**Professor: Sam Plati**

Jordan Ella, 100429623

March 31, 2019

**INTRODUCTION**

Early detection of cancer is critical to increasing treatment success rates but the reality is that doctors do not have the time nor the accuracy required to perform routine screening tests on a large scale. The successful deployment of a model that can reliably detect malignant cells will not only be an improvement over current human ability in both volume of tests and precision, it will free up doctors to focus on other tasks. This will allow for not only a dramatic impact on early cancer screening but free up financial and human resources to positively impact problems outside of the scope of this project. It is for these reasons that Princess Margaret Cancer Centre Clinical Research Unit has requested analysis be performed on their supplied dataset.

**KEY QUESTIONS**

Cancer is most difficult to treat when it is detected late and it can largely be asymptomatic. A predictive model to assess pre-screening test data would greatly improve the rate of successful treatment by enabling earlier detection. This raises a few important questions which will be answered throughout the report and in the conclusion.

* Will this model be successful at detecting early instances of cancer or only once the cancer has developed?
* Which metric(s) can be used to assess the performance of the model?
* Can this model be considered reliable in its detection capability?

**DATA ANALYSIS**

**DATA**

The dataset being analyzed for this project is CancerPreScreenRaw.csv. The data contains 14 independent variables, all measurements of cell structures, and the dependent variable is a binary classification of the cell as cancerous (2) or not (1). The positive results contain only TNM Stage 1 diagnoses usually meaning the cancer is small and has not spread. This last note is significant as it speaks to the first of the key questions. If the model is determined to be accurate, it will fit the criteria of detecting early instances since the data provided is exclusive to early detected cases.

**METHODS**

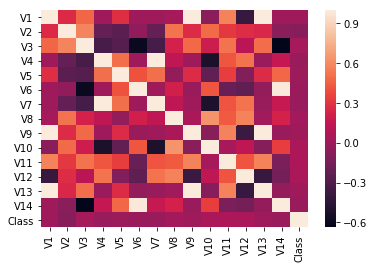
To begin exploratory analysis the dataset was assessed for null values. There were 9 rows which featured null values for each column so the records were simply omit from the dataset.

When assessing the distribution for the independent variables there were some significant outliers observed. While some algorithms might not be affected others are severely impacted by such issues and so the decision to remove rows featuring the outliers was made. These outliers were identified by performing z-score calculations on the dataset and removing those records featuring a z-score of 3 or greater. This purged 88 rows of data that featured values more than 3 standard deviations from the mean. The independent variable class is binary and distributed with an approximately 55/45 ratio.

**DATA ANALYSIS CONTINUED**

**METHODS CONTINUED**

A few columns were identified as strongly correlating with each other, notably V1 and V9, V1 and V13, V4 and V7, V6 and V14, and V9 and V13. These can be observed in the correlation heatmap below, highlighted in a light colour.



**ANALYSIS AND RESULTS**

**TRAINING**

The dataset features almost 15000 rows and so the decision was made to reserve 90% of the data for training and 10% simply for testing. To accommodate some of the binary models explored, the independent variable, Class, which is featured as 1 or 2, was encoded to be 0 or 1, formerly 1 or 2 respectively.

**ANALYSIS AND RESULTS CONTINUED**

**TESTING**

**Random Forest**

While the Random Forest model was not subject to refinement, its initial results leave a lot to be desired. Note the low recall for cancerous cells, far below our target, in this classification report.

RMSE : 0.30543181007227593

Model Score : 0.9067114093959732

[[750 40]

[ 99 601]]

precision recall f1-score support

1 0.88 0.95 0.92 790

2 0.94 0.86 0.90 700

micro avg 0.91 0.91 0.91 1490

macro avg 0.91 0.90 0.91 1490

weighted avg 0.91 0.91 0.91 1490

**k-NearestNeighbors**

Initial investigation with this algorithm yielded less than desirable results but this algorithm benefit significantly from the identification and omission of outliers. Once that was addressed it improved dramatically. Recall, our key metric for evaluation, was 98.4%. This can be observed from the confusion matrix and classification report below.

RMSE : 0.1320971779771627

Model Score : 0.9825503355704698

[[775 15]

[ 11 689]]

precision recall f1-score support

1 0.99 0.98 0.98 790

2 0.98 0.98 0.98 700

micro avg 0.98 0.98 0.98 1490

macro avg 0.98 0.98 0.98 1490

weighted avg 0.98 0.98 0.98 1490

**ANALYSIS AND RESULTS CONTINUED**

**TESTING CONTINUED**

**LightGBM**

The LightGBM model was very accurate as can be observed from the classification report below although it did not reach the same level as the k-NearestNeighbors model. There is a large variety of parameters that could be tuned on this model to improve its performance though so it will certainly be investigated further.

RMSE : 0.18134471605279393

[[763 27]

[ 22 678]]

precision recall f1-score support

1 0.97 0.97 0.97 790

2 0.96 0.97 0.97 700

micro avg 0.97 0.97 0.97 1490

macro avg 0.97 0.97 0.97 1490

weighted avg 0.97 0.97 0.97 1490

**VALIDATION**

Cross-validation was employed to verify that the model was not overfitting or underfitting the data. K-Fold CV, Repeated K-Fold CV, and Shuffle-Split techniques were used. Another popular technique, Leave-Out-One, was omitted due to the computational complexity from such an exhaustive technique on a dataset of this size. The root-mean-squared-error calculations for the cross-validation comparisons can be seen in the table below.

Algorithm Name Original KFcv rKFcv ShuffleSplit

KNeighborsClassifier 0.13 0.14 0.14 0.14

As can be observed, the original model varies only slightly from the other techniques, well within acceptable variance. The RMSE is lower than the cross-validation models indicating that the model might be slightly overfitting but not in any dramatic fashion.

**CONCLUSIONS**

The model will be successful at detecting early instances of cancer as it was trained on data with positives exclusive to early detection. k-Nearest-Neighbors was a very effective algorithm for our problem and was successful in reaching our target recall for cancerous cells with a 98.4% recall rate which we identified as our key metric. This solution is not very computationally demanding for a dataset of this size and is demonstrated through cross-validation to be reliable in its predictive ability.

**APPENDIX**

For all the Python code used in this analysis see the attached HTML file for the project notebook.