Paper PP29

Interactive Real-World Data Analysis Using Machine Learning Algorithms

James J. Kim, Pfizer Inc., Cary, U.S.A.

ABSTRACT

The MDV (Medical Data Vision) database is based on health claims, administrative and DPC (Diagnosis Procedure Combination) data from over 300 Japanese hospitals which is used to investigate and research the Japanese market including pharmacovigilance and epidemiology studies. As of January 2020, it contains approximately 30 million patient records, 1.7 billion disease records and 9.2 billion activity (treatment, intervention, etc.) records.

In RWD analysis, common work is to clarify actual usage of the target drug. For example, we like to know demographics (i.e. age, weight and etc.), frequently used dosage, common tests for diagnosing diseases or concomitant medications. It would be ideal to generate descriptive analysis results identifying the primary factors, yet it is challenging to find common patterns in complex conditions involving multiple diseases, medications and tests.

The emergence of recent machine learning technology and AI, together with the availability of affordable high-performance computing systems have the potential to accelerate and automate the analysis process. We wanted to build an interactive analysis procedure where we can modify different machine learning parameters and algorithms to optimize the output. While both Python (Jupyter Notebook) and R (R Markdown) provide interactive environments for users, we've selected Python for its popularity in data science space and large choices of pre-built packages.

INTRODUCTION

In this pilot study, we extracted two cohort sample datasets from the MDV database between April 1, 2010 to January 31, 2018 – the cohort 1 contained 1,000 patients whom were diagnosed with PAH (Pulmonary Arterial Hypertension) and the cohort 2 contained 10,000 general patients whom were selected randomly. The datasets contained both disease and medical procedure information (numerically coded) but we only used the medical procedure data as the disease data can reveal the patient condition and should be used for validation purpose only.

METHODS

1. Load the datasets.
2. Identify the variables to use for training. Initially we used the datasets as they are and selected Age, Code and Sex as the feature variables (See Figure 1) but this resulted in poor accuracy. With further examination, we identified that they have to be transformed and coded (i.e. one-hot encoding). So, we ended up with 12,450 feature variables (i.e. X) and 1 target variable (i.e. y) matrices. There were 11,000 patients (i.e. 1,000 with PAH and 10,000 general) which were randomly assigned into the training and testing data in the ratios of 0.7 and 0.3. So, we ended up with [7,700, 12450] train matrix and [3,300, 12450] test matrix.
3. We tested several ML algorithms during the process but selected two to show their performance difference in the outcome: Decision Tree Classifier and XGBoost Classifier. They both provide a good starting point and provide feature importance (i.e. which variable contributed the most in the decision process) as being tree boosters as the base algorithm.
4. We produced the following metrics:

    A. Predicted Probability

    B. Confusion Matrix

    C. Classification Report

    D. ROC Curve

RESULTS

Figure 1:
Decision Tree Classifier without the data transformation

```
0.7477152765796784
[[751333 161353]
 [242241 444829]]
                precision    recall  f1-score   support

             0       0.76      0.82      0.79    912686
             1       0.73      0.65      0.69    687070

    micro avg       0.75      0.75      0.75   1599756
    macro avg       0.75      0.74      0.74   1599756
weighted avg       0.75      0.75      0.75   1599756
```

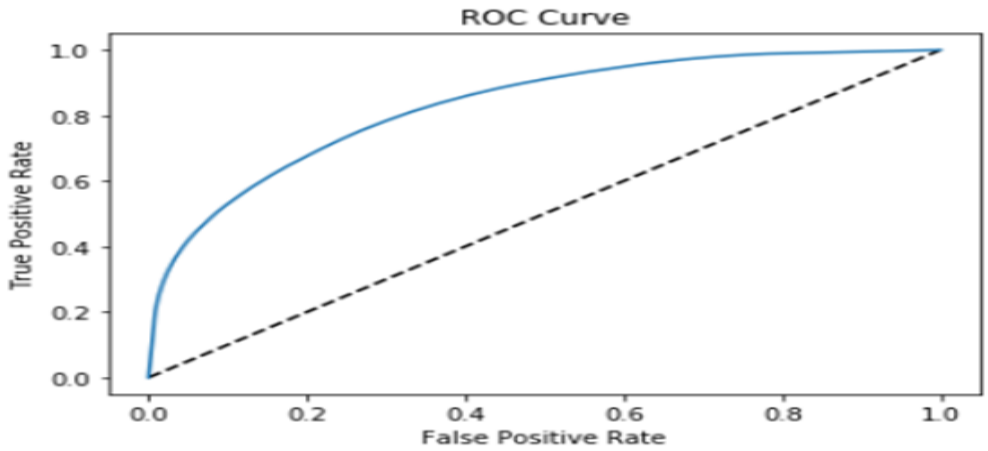


Figure 2:
Decision Tree Classifier with the data transformation

```
0.9881818181818182
[[2994   16]
 [  23  267]]
                precision    recall  f1-score   support

             0       0.99      0.99      0.99      3010
             1       0.94      0.92      0.93       290

    micro avg       0.99      0.99      0.99      3300
    macro avg       0.97      0.96      0.96      3300
weighted avg       0.99      0.99      0.99      3300
```

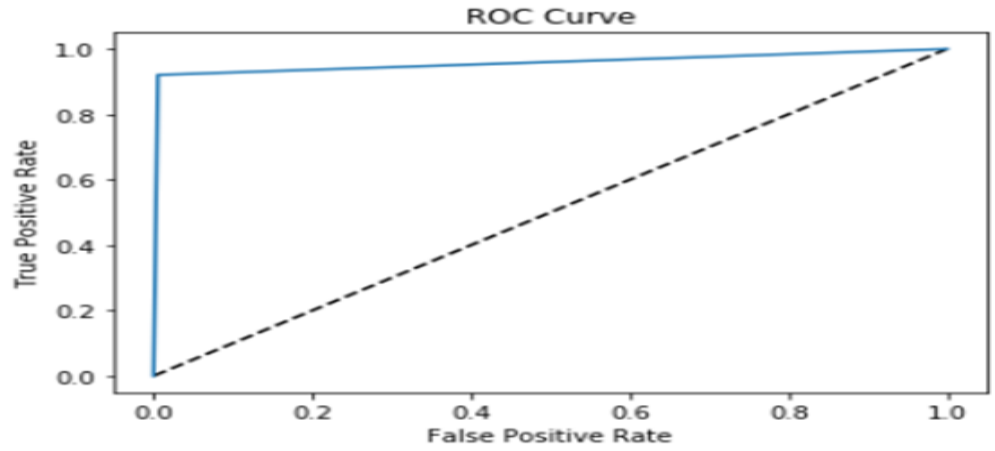


Figure 3:
XGBoost Classifier with the data transformation

```
0.9972727272727273
[[3009    1]
 [   8  282]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3010
           1       1.00      0.97      0.98       290

    micro avg       1.00      1.00      1.00      3300
    macro avg       1.00      0.99      0.99      3300
weighted avg       1.00      1.00      1.00      3300
```

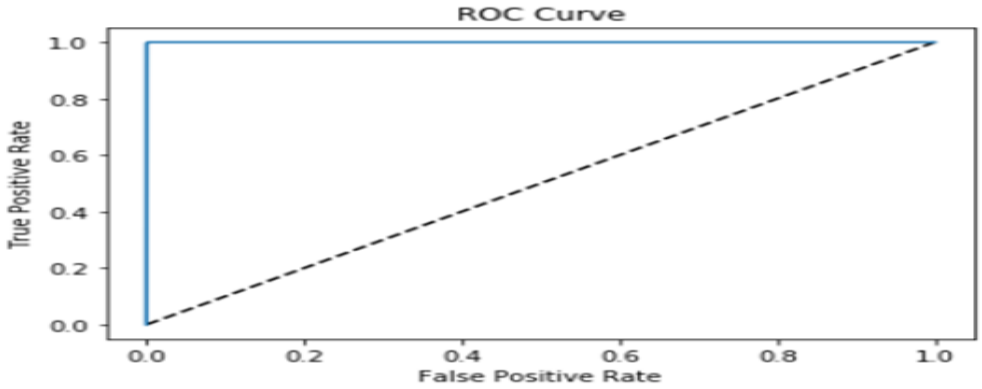
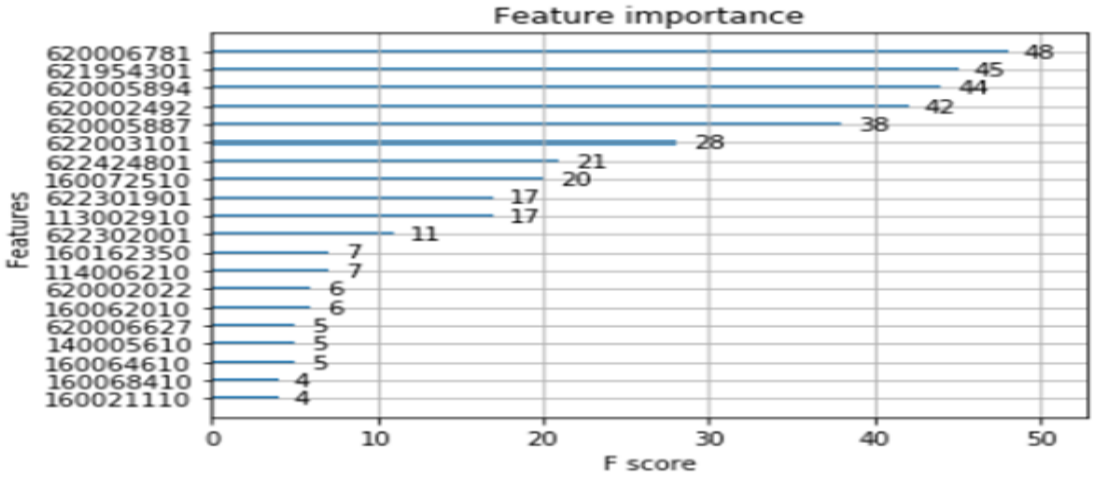


Figure 4:
Feature Importance based on XGBoost Classifier



CONCLUSION

Although Figure 1 appears to generate a reasonable accuracy (74.7 %), this number makes little sense - this is due to the fact that the data was not properly aligned for computation. In Figure 2, it shows that the result significantly improved (98.8 %) after the data transformation while the ML algorithm (Decision Tree) remains the same. Figure 3 shows that changing the algorithm to XGBoost enhances the result accuracy to 99.7 %.

Figure 3's confusion matrix shows that XGBoost identified 282 PAH patients correctly while falsely identified 8 of them as non-PAH patients out of 290 in total. Also, it only identified one patient as PAH patient out of 3010 non-PAH patients, nearly accomplishing 100% accuracy. This is also evident in the ROC curve given.

Figure 4 shows which codes (i.e. medications or procedures) contributed the most in the descending order. I have decoded them few and listed below for your reference:

- ***620006781*** *– Revatio (Sildenafil Citrate – Viagra) is the active ingredient in both branded Viagra and the drug Revatio, which is a lower dose than Viagra and approved to treat pulmonary arterial hypertension*
- ***621954301*** *– Adcirca (Tadalafil – Cialis) is not prescribed to treat erectile dysfunction, despite the fact that it is the same generic drug as Cialis. Adcirca, is prescribed as treatment for PAH. Both Cialis and Adcirca are brand names of the generic drug Tadalafil. Tadalafil is a PDE5 inhibitor.*
- ***620005894*** *– Berasus LA Tablets (Beraprost sodium) – This medicine prevents blood clotting and dilates blood vessels by exerting an action on the prostacyclin receptors on pulmonary vessel smooth muscles and platelets, thus improving the blood flow.*
- ***620002492*** *– TRACLEER is a prescription medicine indicated for patients with certain types of PAH.*

Figure 5 shows the tree plot based on the Figure 3 and 4 (XGBoost model).

The sample data used in this study was only a small fraction compared to the entire MDV database and the algorithms used were basic; yet we were able to achieve an extremely efficient and accurate machine learning model. Imagine what we can do with more data and advanced algorithms such as deep learning (e.g. Tensorflow and Keras). It will open up new possibilities of solving complex RWD problems and answering questions which we never thought of asking.

I plan to work with the Japan RWD team again in 2020 to improve and optimize the existing models for further investigation – possibly using different sample sizes and/or databases.

REFERENCES

Python - https://www.python.org/

NumPy - http://www.numpy.org/

Pandas - https://pandas.pydata.org/

Scikit-learn - https://scikit-learn.org/stable/index.html

XGBoost - https://github.com/dmlc/xgboost

Anaconda - https://www.anaconda.com/

Jupyter - https://jupyter.org/

Tensorflow - https://www.tensorflow.org/

Keras - https://keras.io/

ACKNOWLEDGMENTS

There are many colleagues to mention but special thanks to:

Stuart Pearce and Patti Compton - the management support.

Mikio Hayashi, Shintaro Hiro and Yoichi Ii – the Japan RWD team.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

James J. Kim
Pfizer Inc.
781-429-5109
james.j.kim@pfizer.com
linkedin.com/in/kimjj/