09-615 Computational Modeling in Science

Final Project

12/09/2024

Anagha Anil, Anna Bolling, Andrew Lutsky, Xiaolei (Brian) Zhang

anaghaan@andrew.cmu.edu, ambollin@andrew.cmu.edu, alutsky@andrew.cmu.edu, xiaoleiz@andrew.cmu.edu

The objective of our project was to analyze the consolidated BRFSS 2015 (Behavioral Risk Factor Surveillance System) dataset provided by the CDC, which was based on a survey of over 400,000 Americans about their behaviors, health conditions, and whether they have been diagnosed with diabetes/prediabetes. We aimed to analyze these health indicators to determine if the survey is an accurate assessment of diabetes health risk, and if certain attributes are more or less indicative of diabetes. To do this, we employed an exploratory data analysis and feature selection to narrow down the dataset, and used logistic regression, decision tree, random forest, and XGBoost classification models to assess the usefulness of the dataset.

1. Data Loading and Preprocessing

The initial dataset was retrieved from a Kaggle competition (https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data), and consisted of a basic CSV file with 253681 observations and 22 health indicators (features) such as BMI, Age, Education Level, and Income. We imported key libraries like pandas, numpy, scikit-learn and matplotlib and checked how many NaN or missing values were present in the dataset. From this we learned that the dataset was well constructed with no missing values; however, we performed a preprocessing step by converting all Diabetes_binary values (target variable) from 0, 1, 2 to just 0, 1. A Diabetes_binary value of 0 indicates no diabetes diagnosis, 1 prediabetes, and 2 diabetes. By doing this, we were able to use binary classification models such as logistic regression in our model exploration steps.

2. Feature Selection and Scaling

From an initial UMAP and PCA of the data, we learned that the dataset was slightly clustered into the no diabetes and diabetes groups, although a fair amount of scatter and noise remained. We made an initial plot of variance explained vs number of components used for the full dataset, and found that 90% variance explained was achieved in about 18 features (with 22 features being present total). We then used Logistic Regression and XGBoost importance scores to visualize the importance of the 22 features in the dataset; we found that three features were much more important than the others for explaining variance in the dataset: HighBP, GenHealth, and HighCholesterol. To reduce dimensionality and focus on these impactful variables in our model

exploration step, we selected the top 10 most important features deemed by XGBoost and created a trimmed dataset. Using these selected features, visualizing the variance explained vs components used resulted in 90% variance explained achieved in just 8 features.

Following this, we made a quick pie chart of the amount of diabetes vs no diabetes classes in the full dataset and found that there existed significant class imbalance (No Diabetes – 86%, Diabetes – 14%). Initial runs of our models yielded low F1-scores of around 0.2 to 0.3, which could be explained by this class imbalance. To address this, we took the trimmed dataset and randomly selected observations from both No Diabetes and Diabetes classes to create a trimmed and balanced dataset.

3. Model Training and Cross-Validation

We used scikit-learn's Pipeline package to set up an efficient pipeline for testing and model training. We decided to test Logistic Regression first as a linear model, then Decision Tree as a nonlinear model, then Random Forest, XGBoost, and Gradient Boosting as more powerful ensemble models. This approach of both linear and nonlinear models seemed appropriate as some health indicators in the dataset such as BMI have been shown to have nonlinear associations with diabetes, while others such as income have more linear associations (Zaccardi et al., 2017). The trimmed and balanced dataset was separated into an 80:20 training-testing split, with StandardScaler being used to normalize continuous variables such as BMI and Age. Cross-validation with 3 folds was performed to fairly evaluate model stability.

From initial runs of our pipeline on the imbalanced dataset, we found that ensemble models such as XGBoost achieved the highest accuracy at 86.7%, while having very low F1 Scores at 25.7%. Running the pipeline on the trimmed and balanced dataset revealed that XGBoost still was able to achieve the highest accuracy of the models at 75.1% with an F1 Score of 76.1%, while also having consistent accuracy across cross-validations (Mean Accuracy = 0.7520). Interestingly, the worst performing model on the trimmed and balanced dataset was Decision Tree, which had an accuracy of 66.8%; this was followed by Random Forest with an accuracy of 70.6%, then Logistic Regression with an accuracy of 74.4%. While we expected the ensemble models to outperform the simple linear/nonlinear models, this result could be a result of model overfitting.

4. Hyperparameter Tuning

Lastly, we attempted to maximize the performance of our models by using Optuna to find the best set of hyperparameters for each model using Random Search. We explored hyperparameters such as max_depth and min_samples_split for Decision Tree, number_estimators for Random Forest, and learning_rate for XGBoost. Following hyperparameter tuning, we found that XGBoost was still the best performing model with the highest accuracy (75.2%) and F1-Score (76.1%). Other models such as Decision Tree improved in accuracy substantially following hyperparameter tuning, being able to improve nearly 8 percentage points in both accuracy and F1 score.

5. Conclusions

From our analysis of the BRFSS 2015 diabetes health indicators dataset, we learned that addressing class imbalance and performing feature selection is key to maximizing the performance of model training and tuning. We also determined that for this particular dataset, XGBoost is the best performing model, being able to accurately determine diabetes diagnoses given health indicator data for approximately 75% of patients. We found that the most important health indicators were high blood pressure, high cholesterol, and self-assessed general health, which could be key factors to focus on for future health risk analysis surveys.

References

Burrows, N. R., Hora, I., Geiss, L. S., Gregg, E. W., & Albright, A. (2017). Incidence of End-Stage Renal Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes - United States and Puerto Rico, 2000-2014. *MMWR. Morbidity and mortality weekly report*, *66*(43), 1165–1170. https://doi.org/10.15585/mmwr.mm6643a2

Zaccardi, F., Dhalwani, N. N., Papamargaritis, D., Webb, D. R., Murphy, G. J., Davies, M. J., & Khunti, K. (2017). Nonlinear association of BMI with all-cause and cardiovascular mortality in type 2 diabetes mellitus: a systematic review and meta-analysis of 414,587 participants in prospective studies. *Diabetologia*, *60*(2), 240–248. https://doi.org/10.1007/s00125-016-4162-6