DETECTION OF DIABETES AMONG PATIENTS WITH VARIOUS RISK FACTORS

By: Rohit Varanasy, Gayatri Chabra, Srinidhi Manikantan

INTRODUCTION

Diabetes, a prevalent chronic disease in the United States, significantly impacts health and the economy.With 34.2 million diagnosed cases and 88 million prediabetic individuals (CDC's 2018 data), the lack of awareness among 1 in 5 diabetics and 8 in 10 pre diabetics is concerning. As diabetes prevalence rises, prioritizing early detection and targeted interventions becomes crucial.

PROJECT AIMS

Our project is intricately woven with a dual focus: investigating the likelihood of diabetes through the analysis of diverse risk factors and emphasizing the practical impact of our predictive models. Unlike traditional methods that heavily lean on approaches like logistic regression, our methodological perspective embraces the utilization of more robust machine learning techniques, notably ensemble methods, recognized for their efficacy in capturing complex relationships. Our overarching goal is to attain practical significance by emphasizing the early identification of high-risk individuals and contributing to targeted strategies for reducing diabetes prevalence. Additionally, we aim to advance the field by underlining the superiority of ensemble methods and pinpointing specific factors strongly linked to diabetes.

DATASET AND EXPLORATORY DATA ANALYSIS

The dataset was obtained from the Behavioral Risk Factor Surveillance Survey (BRFSS), which is an annual survey conducted by CDC to collect health-specific data. We have chosen the 2015 dataset which has 441k responses and 330 features. Given that the dataset contains generic health information (i.e. not specific to Diabetes), we did feature selection by choosing the top 25 features that were important in influencing diabetes based on research. Feature selection is important in reducing the computational time and improving the model performance.

Data cleaning is an important step in Exploratory Data Analysis as it helps ensure that the data is of the highest quality and improves the modeling process. We removed duplicate values, invalid responses like “Don’t Know/Refused”, and missing data as part of data cleaning. The intention was to preserve as much of the data as possible without compromising on its quality. This process significantly reduced the number of observations by half (i.e. from about 440k to 230k), which could be attributed to the fact that the dataset was based on telephone survey results which is likely prone to more errors. Due diligence was also done to make sure that there were no (i) outliers, (ii) variables with no unique values, (iii) low variance variables, (iv) highly correlated variables, and (v) erroneous data, as all this could affect our analysis significantly.

MACHINE LEARNING MODELS

Initial observations prompted an investigation into the prevalent use of simple models, particularly logistic regression, in the healthcare industry. Logistic regression without penalty served as our baseline, exhibiting an impressive F1 score of 0.794 within a mere 30 seconds. Recognizing the dataset's significant class imbalance, we adopted F1 as our metric for model evaluation, steering away from accuracy to avoid bias. Subsequent models, including penalized logistic regressions (L1 and L2), elastic net, and K-Nearest Neighbors, showcased F1 scores in the range of 0.786 to 0.797, each with varying computational times. Transitioning to ensemble methods, we delved into bagging, boosting, and stacking techniques. XGBoost emerged as the standout performer, yielding an F1 score of 0.801, surpassing other models like random forest. Stacking models, particularly XGBoost and LightGBM, maintained a commendable F1 score of 0.797 while demonstrating efficiency, taking only 13 minutes to run.

Our project revealed crucial insights into the data science pipeline, emphasizing that model complexity doesn't consistently equate to superior performance. Despite being more computationally demanding, ensemble methods did not substantially surpass simpler models. The analysis underscored challenges in healthcare data, marked by irregularities, especially in self-reported or cross-sectional datasets. Future directions include addressing class imbalance, refining precision with concise BRFSS questions, and augmenting predictive power through diverse data sources.

Feature importance analysis, conducted using Sci-kit Learn, pinpointed high blood pressure as the most influential factor in determining a patient's diabetic stage. This outcome underscores the critical role of specific health indicators in predictive modeling.

We also acknowledge the data limitations, such as the cross-sectional nature of the BRFSS data and potential recall bias. This highlights areas for refining our modeling approaches. Despite challenges, our project provides valuable insights into diabetes prediction, paving the way for further advancements in healthcare modeling.

CONCLUSION

In conclusion, our project delved into the multifaceted landscape of diabetes. We highlighted the dual focus of our investigation: understanding the likelihood of diabetes through diverse risk factors and emphasizing the practical impact of predictive models. We sought to develop predictive models using advanced machine learning techniques, departing from conventional methods. While simple models like logistic regression demonstrated efficacy, ensemble methods such as XGBoost and LightGBM exhibited superior performance. While acknowledging the challenges inherent in healthcare data, our project provides valuable insights and underscores the dynamic interplay between data science and healthcare, offering a foundation for future exploration and refinement in the domain of diabetes prediction.