Final Project Report

BAX 423: Big Data Analytics

Andrew Hamaty, Kshitij Karan, Sungho Lee, Namuun Boldbaatar

June 9, 2023

**Business Objective**

The business objective is to create a predictive model that assesses the risk of diabetes using data gathered from the Behavioral Risk Factor Surveillance System (BRFSS) survey. Our primary goal is to enhance the identification of individuals prone to diabetes and enable the implementation of effective preventive measures. By designing an advanced predictive model, we aim to aid public health officials and medical experts with a valuable tool for accurately identifying individuals who are at risk of developing diabetes.

**Key Actionable Business Initiative**

The first business initiative is to develop a targeted marketing strategy to identify key demographics and users that are most likely to adopt our platform. From there, we would like to pursue three different revenue streams. The first revenue initiative is to develop a licensing subscription model. We hope to offer healthcare institutions, clinics, and research facilities the opportunity to license our platform and models on a subscription basis. Next, we wish to pursue integration partnerships with bioengineering companies to integrate our models with other software and hardware, enabling seamless integration of our product with existing and possibly new platforms.

**Metrics of Success & KPIs**

In order to assess the effectiveness of our business objective, we will evaluate the precision of the developed predictive model in identifying individuals with a higher likelihood of developing diabetes. The evaluation will be based on essential metrics such as sensitivity, F1-Score, and overall accuracy of the model. However, we will emphasize the significance of sensitivity, F1-Score, and positive predictive value as the primary metrics for gauging success.

**Role of Analytics**

Analytics plays a key role in our effort to build a predictive model for assessing diabetes risk. It aids us in identifying and incorporating crucial variables associated with demographics, lifestyle, and medical factors. Leveraging analytics enables us to gain a comprehensive understanding of the interrelationships between these variables and the risk of diabetes. This understanding empowers us to identify populations at high risk and devise effective preventive strategies. It also allows us to gauge the accuracy of the developed predictive model in identifying individuals who are susceptible to diabetes. Additionally, in the future, analytics enables us to measure key performance indicators such as user adoption, conversion rates, and client satisfaction, providing actionable insights to evaluate the success of the initiative, make informed decisions, and set benchmarks for the future.

**Analytics Methodology & Description of Data**

Our analytics methodology leverages a dataset available on Kaggle, specifically the 2015 BRFSS survey, to enhance our understanding of diabetes risk. This dataset encompasses responses from over 400,000 Americans, offering valuable insights into health-related risk behaviors, chronic health conditions, and the utilization of preventative services. With 22 features encompassing demographic information, medical history, lifestyle factors, and health-related behaviors, we have a robust foundation for our target outcome: developing a predictive model for diabetes risk.

**Type of Analytics and Methodology**

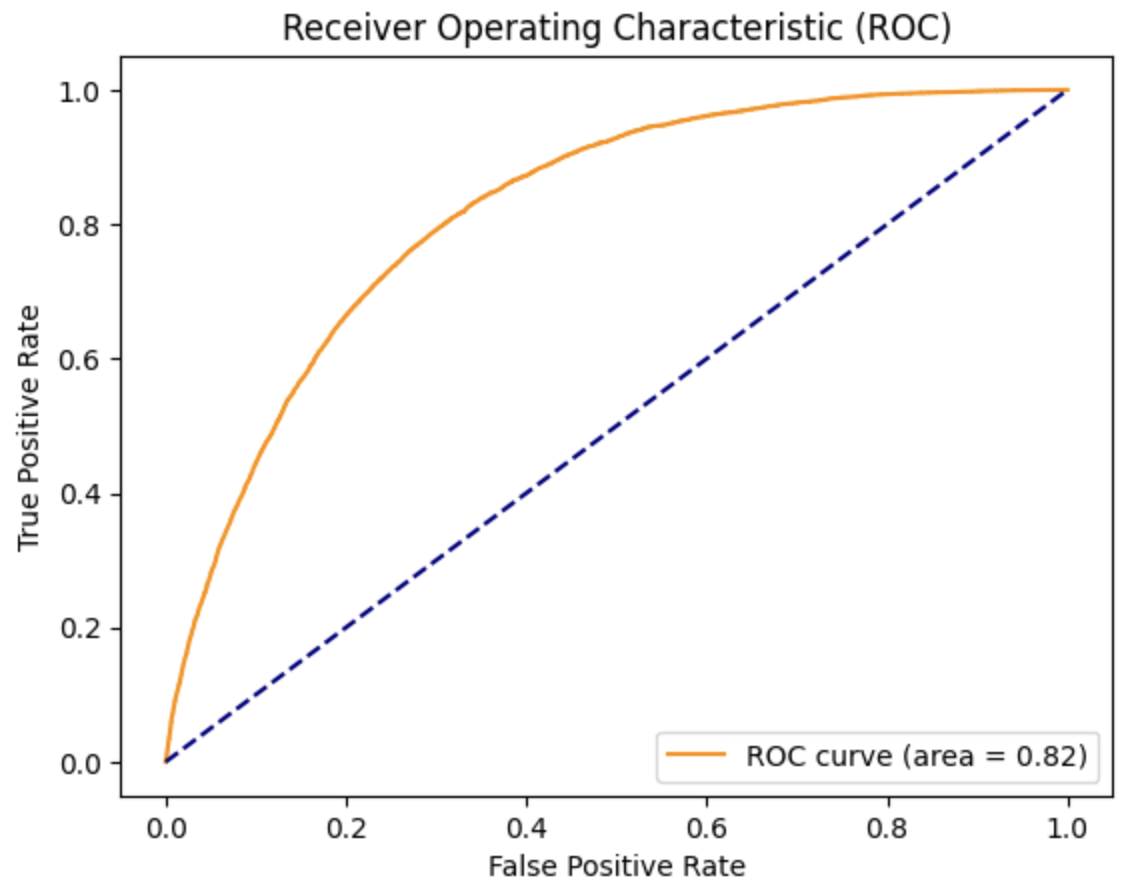
To accomplish this goal, we will employ a combination of descriptive, predictive, and prescriptive analytics. Descriptive techniques were used to thoroughly explore and document variations present within the data. Through this process, we identified key trends, patterns, and relationships that will serve as valuable input for developing our predictive model for diabetes risk. Additionally, we utilized predictive analytics to devise our predictive, logistic regression model, as well as identify features that increase the likelihood of diabetes. Lastly, we used prescriptive analytics to draw causal inferences from the identified features using matching algorithms. By utilizing analytics, we aim to craft a precise and reliable predictive model capable of identifying individuals at risk of developing diabetes. This, in turn, will enable timely interventions and implementation of preventative measures, contributing to improved healthcare outcomes.

The logistic regression model will be used to obtain probability scores for diabetes. These probabilities will serve as “diabetes risk scores” for our further analysis. Moreover, from the identified important features, we will select the feature that can resemble a “treatment” in a randomized control trial.

Since the data is observational, we cannot run a Randomized Controlled Trial (RCT) to draw causal relationships. However, we will be using propensity matching using the nearest neighbor algorithm to create matched pairs. This will serve as a quasi-experiment for our analysis. Using these matched pairs of treatment and control units, we will evaluate the Average treatment for treated (ATT), Average treatment for untreated (ATU), and the Average Treatment Effect (ATE).

**Main Analytics Results**

Our logistic regression model achieved an accuracy of 86.3%, a precision of 83%, a recall of 86.3%, and an F1 score of 84.6%. Additionally, we achieved an AUC of 82%. When attempting to fit the model into each age group, we found that the model performed better with some age groups than others. The model is most accurate at predicting the likelihood of diabetes for younger age groups, more specifically those aged 18-49. Model accuracy slightly dips when attempting to create predictions for ages 50+, however, remains over 80%. The age group that was most difficult for the model to predict was those aged 70-74.



As per the model, we identified four important features for evaluating the likelihood of diabetes: cholesterol awareness, heavy alcohol consumption, high blood pressure, and high cholesterol levels. Out of these 4 features, 3 of them are related to health status while heavy alcohol consumption can be attributed to lifestyle.

In prescriptive analytics, we used propensity-matching scores to create a quasi-experiment. For the match-paired tests, we used individuals with high levels of alcohol consumption as the treatment and the probability of diabetes as the outcome. We discovered that there was a significant relationship between heavy alcohol consumption and diabetes. Individuals with heavy alcohol consumption were much more likely to have diabetes.

The “treatment” group was heavy alcohol consumption. The causal impact of heavy alcohol consumption on diabetes risk was found to be 42.7%. This is a relatively large figure, therefore our recommendation would be to reduce alcohol consumption to reduce the risk for diabetes.

**Conclusion**

In conclusion, our business objective was to create a predictive model that assesses the risk of diabetes using data from the BRFSS survey. The primary goal was to enhance the identification of individuals prone to diabetes and enable the implementation of effective preventive measures. Our proposed targeted marketing strategy outlined revenue streams, including a licensing subscription model and integration partnerships. The success metrics and key performance indicators focused on evaluating the precision of the predictive model, with sensitivity, F1-Score, and positive predictive value highlighted as primary metrics. Analytics played a crucial role in identifying important variables, understanding their interrelationships, and assessing the accuracy of the model. Our analytics methodology encompassed a comprehensive blend of descriptive, predictive, and prescriptive analytics techniques, allowing us to derive meaningful insights, develop accurate predictions, and provide actionable recommendations based on our findings. The logistic regression model achieved promising results, with high accuracy, precision, recall, and F1 score, as well as a respectable AUC. Important features associated with diabetes risk were identified, including cholesterol awareness, heavy alcohol consumption, high blood pressure, and high cholesterol levels. Additionally, the results of the quasi-experiment and propensity-matching revealed a significant relationship between heavy alcohol consumption and diabetes. The causal impact of heavy alcohol consumption on diabetes risk was found to be substantial, suggesting a recommendation to reduce alcohol consumption to mitigate the risk of diabetes.