Fall 2021 DSGA – 1001 Introduction to Data Science

Capstone Project – Team Insulin

Diabetes - A Deep Dive

Authors

- Varsha Vattikonda - vv2116
- Tharangini Sankarnarayanan - ts4180
- Shriya Murthy Akella - sa6523
- Shambhavi Sachin Rege - sr6172

Details

- Git Project
- Kaggle Data
- Kaggle Notebook

Introduction

Diabetes was the ninth leading cause of death in 2018, accounting for an estimated 1.5 million deaths [1]. Diabetes is a common disease that affects blood glucose regulation. It can harm other vital organs in the body if not treated promptly. According to the Centers for Disease Control and Prevention (CDC), 34.2 million Americans had diabetes in 2018, and 88 million had prediabetes [2]. Diabetes is a chronic condition that occurs when the body cannot produce insulin (insulin is the hormone that regulates blood sugar) or when the body's insulin is not processed properly, leading to dangerously high blood glucose levels. Diabetes is incurable, but it is manageable. Diabetes over time leads to serious damage to the eyes, nerves, and blood vessels, causing heart attacks, strokes, and kidney failures [3]. We aimed to develop predictive models to identify risk factors for diabetes, which could help facilitate early diagnosis and intervention and reduce medical costs.

There are many existing models for the prediction of diabetes. However, due to the complexity of diabetes's causality, the prediction performance (particularly sensitivity) of models based on survey data needs to be improved. Furthermore, while many diabetes risk factors, such as obesity and age, have been identified, others have yet to be identified. We use the Behavioral Risk Factor Surveillance System (BRFSS) [4], an annual health-related telephone survey conducted by the Centers for Disease Control and Prevention (CDC) to identify various risk factors. To investigate the statistical significance of our hypotheses, we conduct hypothesis testing. The goal of our study was to develop diabetes predictive models using 2014 BRFSS data and machine learning techniques of Decision Trees, Random Forest, XGBoost, and Convolutional Neural networks.

Data Preparation

This original survey data has 330 features and responses from 441,455 people. The dataset for this project was obtained from Kaggle (Diabetes Health Indicators Dataset) [5], and it is a cleaned dataset of 70,692 survey responses to the 2015 CDC survey. This project's dataset contains 21 features and 253680 participants. The problem as one can guess is an imbalanced classification problem. The dataset includes approximately half of the respondents who do not have diabetes and the other half who do have diabetes or prediabetes.

Every row of the dataset represents a survey response. The columns of the dataset represent the features/parameters that contribute to predicting the risk of diabetes in the dataset are as follows:

1. High blood pressure: *1* indicates hypertension, while 0 indicates normal blood pressure.
2. HighChol: A value of *1* indicates high cholesterol, while *0* indicates low/normal cholesterol.

3. CholCheck: *1* indicates that the participant has had a cholesterol test within the last five years, while a *0* indicates that they have not.
4. BMI stands for Body Mass Index.
5. Smoker: *1* indicates that the participant has smoked at least 100 cigarettes in his entire life. *0* indicates that the participant has smoked less than 100 cigarettes in his entire lifetime.
6. Stroke: *1* indicates that the participants had a stroke and 0 indicates otherwise.
7. HeartDiseaseorAttack: Participant has/had a myocardial infarction or heart disease is indicated by 1 and otherwise by 0.
8. PhysActivity: Participant's involvement in any kind of physical activity apart from a job in the past 30 days is indicated by 1 otherwise *0.*
9. Fruits: *1* is indicative of the participants' fruit consumption being 1 or more times per day, whereas 0 indicates otherwise.
10. Veggies: *1* is indicative of the participants' vegetable consumption being 1 or more times per day, whereas 0 indicates otherwise.
11. HvyAlcoholConsump: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) are indicated by 1 and non-heavy drinkers are indicated by 0.
12. AnyHealthCare: The participants having health care coverage or health care insurance or prepaid plans are indicated by 1 and the ones without these are indicated by 0.
13. NoDocbcCost: *1* indicates that the participant had to see a doctor in the past 12 months but could not because of the cost. *0* indicates otherwise.
14. GenHlth: Self-rating by the participant of his/her general health on a scale of 5.
15. MenHlth: Poor mental health (stress, depression, and other mental health issues) days measured in some days in the past 30 days.
16. PhysHlth: Poor physical health (injuries and other physical health issues) days measured in the number of days in the past 30 days.
17. DiffWalk: *1* indicates that the participant has difficulty in climbing stairs and walking, 0 indicates otherwise.
18. Sex: *1* indicates a female participant, 0 indicates a male participant.
19. Age: *1* indicates age range 18-24, 2 indicates age range 25-29, 3 indicates age range 30-34, 4 indicates age range 35-39, 5 indicates age range 40-44, 6 indicates age range 45-49, 7 indicates age range 50-54, 8 indicates age range 55-59, 9 indicates age range 60-64, 10 indicates age range 65-69, 11 indicates age range 70-74, 12 indicates age range 75-79, 13 indicates age range 80 or older.
20. Education: A participant who has never attended school or only kindergarten is indicated by 1, a participant who has attended grade 1 through 8 (Elementary) is indicated by 2, who has attended grade 9 through 11 (Some high schools) is indicated by *3*, grade 12 or GED (High school graduate) is indicated by 4, *5* indicates college 1 year to 3 years (Some college or technical school), 6 indicates college 4 years or more (College graduate)
21. Income: A participant's income on a *1-8* scale. *1* indicates less than $10,000 and 8 indicates $75,000 or more.

Q1: Is the CDC self-assessment questionnaire valid?

The CDC has a self-assessment form [6] where they ask a series of questions to help anyone determine they COULD be diabetic. If the candidate is in the safe zone for ALL the selected features, the candidate is most likely not diabetic. Using the data, we have; we want to validate this hypothesis.
The features used by the CDC questionnaire are:

1. Age - 'Age' (>=40)?
2. Status of General Health - 'GenHlth'
3. Physical Activity - 'PhysActivity'
4. Body Mass Index - 'BMI'
5. Smoker or Non-Smoker - 'Smoker'
6. ***Family History - We do not have this data and hence we are excluding this feature from our analysis

In terms of our problem, the problem statement is, 'If an individual is in a safe zone for all the features by CDC, is it enough to conclude that the individual is not diabetic'?

The null and alternative hypotheses are defined as:
H0: P (diabetes | given safely for select features) = 0
H1: P (diabetes | given safely for select features) > 0

We choose the Chi-Square Test, as we want to compare the observed and the expected frequencies of the set of individuals who are diabetic and are safe for the selected features (the selected features are categorical). All the selected features are Boolean and hence we have engineered a new feature called safe which is a sum of all features above. This way when an individual is safe for a select feature the new column would be 0, else it would be 1. Chi-Square test gives p-values in the order of -100 and hence we can choose to conclude that the P (diabetes | safe) >0 i.e., the alternate hypothesis is true, and that the data does not support the CDC version of declaring any individual diabetes-free.

Q2: What risk factors are most predictive of diabetes risk? And do they differ for different groups of individuals?

Background

The goal is to select the most relevant factors from the available data that are significant in predicting diabetes risk for any given individual. The first question that arises in this regard is whether these characteristics differ for different groups of individuals, i.e., whether the key factors differ for smoker's v/s non-smokers; younger v/s older; female v/s male.

Problem Statement

Divide the data into possible different groups; Build a different classification model using Decision Trees for each group to compare the feature importance for prediction across different groups i.e., build a model each for Smokers vs Non-smokers and compare the feature importance of both the models. If the feature importance is similar, we might conclude that the key features are the same for different groups of individuals. We repeat the test for different groups like

- Smokers' v/s Non-smokers
- Young, aged v/s old aged
- Female v/s Male

Feature Importance

Consider, we build two decision trees for Female's v/s Males. The resultant feature importance of both was very similar. Checking for equality between the two to confirm feature importance for the gender feature, we check for the feature importance are equal. For this task, we use an independent t-test to check whether the feature importance values we got in both scenarios are equal.

Conclusion

We obtain a p-value of around 0.999, supporting that the data in both the scenarios are equal and so do not require two different classification models for the groups. We get the same conclusion for

- Smokers' v/s Non-smokers
- Younger v/s older
- Female v/s Male

Building into the idea to evaluate the key factors influencing the risk of diabetes, multiple Random Forest Classification models using different hyperparameters are used, and then take the averages of feature importance as indicated in the Fig 1.

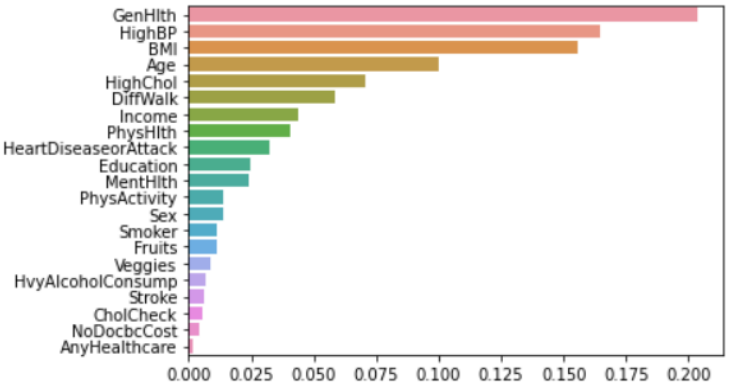

Fig 1: Feature Importance

Modeling & Evaluation

Q3: What is the best fit to predict the risk of diabetes?

Data Preparation

The data as described above is an imbalanced classification problem. So, there is a need for oversampling the data for training. We have used the library imblearn [7] for oversampling the data for the balance of majority and minority classes.

As we do not have train and test data separately, we will be splitting the available data into train and test (⅔ is train data and ⅓ is test data). We perform cross-validation to perform hyper-parameter tuning.

Feature Engineering

To check if we need to choose all the features, we have used the following techniques to try and eliminate the features that are very closely related

- Chi Square Test
- Mutual Information Score.

Since the data is mostly categorical, we did not want to use correlation as the similarity indicator. During the analysis we realized that Chi-Square can only help us understand if the features are very similar but cannot really quantify the similarity. Hence, we have also included Mutual Information Score. But the result was that none of the features were very highly dependent from either of these scores.

Dimensionality reduction

We have chosen PCA for our dimensionality reduction with explained variance as the criterion. We have around 16/21 features explaining around 85% of the variance in the data. And again, since PCA is all dependent on the covariance matrix, is this dimensionality reduction suitable is one of the hyper parameters tuning we are going to do.

Evaluation Metrics

Given that the data is an imbalanced classification problem, using Accuracy as an evaluation metric is not helpful. Hence, we use **F1-Score** (the combination of Precision and Recall) as our primary evaluation metric.

Baseline Model

We have used the package sklearn's **Dummy Classifier** to implement the baseline model. This is a classifier that makes predictions using simple rules. We have used the dummy classifier with strategy as **most_frequent** which implies that it simply predicts the majority class for any given data point irrespective of the data. This model has the following metrics:

	Accuracy	**Precision**	**Recall**	**F1 Score**
Train	0.86	0	0	0
Test	0.86	0	0	0

Table 1: Metrics of Dummy Classifier

Observations

We follow few different options to pursue i.e.

1. Balance the imbalanced classification problem v/s NOT
2. With PCA v/s without PCA
3. Feature selection (using average feature importance from Decision Trees) and then use only the top 80% important features

Given that most of the data is categorical, looking at covariance or performing PCA was our initial intuition, which proved to be correct with the following results (models were created with no hyperparameter tuning to assess the impact of various techniques).

F1-Score	Data as is in Train/Test	PCA	Oversampled	Top 10 Features
Logistic Regression	0.24 / 0.24	0.44 / 0.45	0.45 / 0.46	
Decision Tree	0.31 / 0.98	0.29 / 0.98	0.4 / 0.98	0.3 / 0.98
Random Forest	0.24 / 0.98	0.31 / 0.98	0.4 / 0.98	0.34 / 0.98
XGBoost	0.40 / 0.45	0.1 / 0.2	0.45 / 0.53	0.41 / 0.68

Table 2: Metrics of Machine Learning Classifiers

Hyperparameter Tuning

From the above results we have decided to run hyperparameter tuning in two ways:

- Oversampled data with all the features
- Oversampled data with 80% of the most important features

Since performing cross-validation on the already oversampled data could give very skewed biased results, we must follow the steps in the below specified order:

- Split data into Train data and Test data
- Split Train data into Train data and Cross-Validation data
- Oversample training Data
- Hyper Parameter tuning on training data
- Validation using Cross-validation data
- A final test on testing data

We perform the tuning focusing on the following models and the results are shown in Table 3.

Model	Parameters	Average F1 Score
Random Forest	n_estimators: range (50,200,50) Min_samples_split: range (20,60,10) Min_samples_leaf: range (10,50,10)	0.8
XGBoost	Col_sample_by_tree: range (0.5,0.9) Sub_sample: range (0.5,0.9)	0.8
Multi-Layer Perceptron	Hidden Layer 1 - Neurons: range (1,6) Hidden Layer 2 - Neurons: range (1,6)	0.9

Table 3: Tuning hyperparameters of classifiers

Results

The best fit model is the Multi-Layer Perceptron with the following parameters, resulting in an F1 Score of 0.89 used on test data:

- Learning Rate (alpha) = 1e-5
- Two hidden layers with neurons as (4, 1)

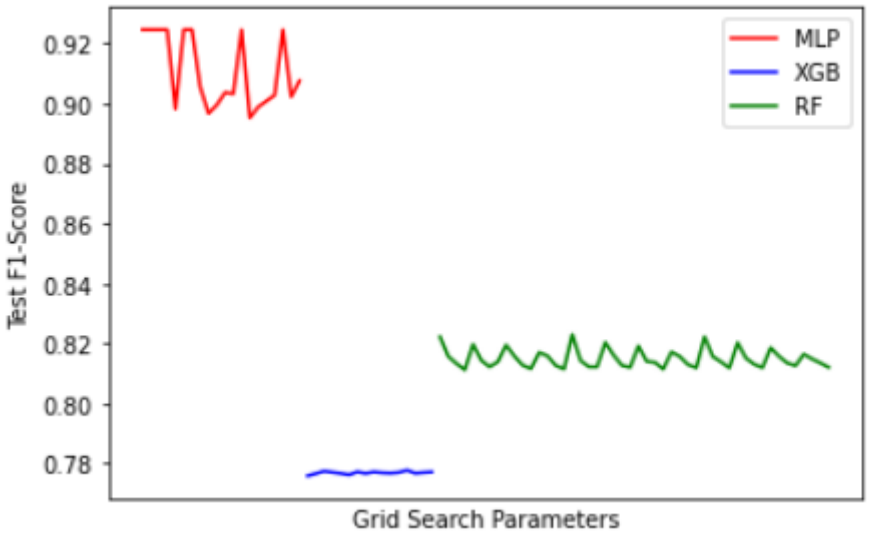

Fig 2: F-1 score of classifiers

Conclusion

Diabetes is a very common problem caused by a variety of factors, so we were eager to see if we could achieve predictability using all the tools provided by the DS-GA 1001 course. The course provided us with so many tools for cross sectioning the data that we realized we had a plethora of options, and it was all about weighing the pros and cons of each tool. From hypothesis testing (which appears to be a very simple concept that comes in very handy) to dimensionality reduction to Neural networks, everything appears to have come together with this one project to establish a good predictive model.
However, the available data has its limitations. With that data, the prediction model could have been even better. More factors like the following could have helped the model be more accurate.

- Choice of Cuisines
- Family history
- Have the changes in BMI over time of an individual

Diving into types of diabetes and their respective factors can also produce more detailed observations and ideas about the various risks in predicting the possibility of diabetes.

References

[1] World Health Organization
[2] Statistics about Diabetes
[3] Scikit Learn library, Python
[4] Behavioral Risk Factor Surveillance System (BRFSS), Center for Disease Control and Prevention
[5] Diabetes Health Indicators Dataset, Kaggle
[6] Self-Assessment Form for Diabetes Center for Disease Control and Prevention
[7] Imblearn library, Python