**Detecting Diabetes**

# Assisting staff with identifying at-risk or positive patients.

**Abstract**

My client is among the largest healthcare network providers in their region. Overfilled beds, understaffed and overworked hospital employees, and undetected illness can be mitigated with database efficiency. We will create a model that assists their practitioners with identifying patients at risk for diabetes, of the most expensive & debilitating conditions impacting Americans.

**Design**

The database provided by Kaggle was downloaded, cleaned, and massaged to create two binary files meant to detect 1) pre-diabetics, and 2) full but undiagnosed diabetes. Non-binApAary fields were scaled. Data was further split into feature-engineering and non-FE groups to determine which were more impactful. Several classifiers were tested on these subgroups, including KNN-neighbors, logistic regression, decision trees, extra trees, and random forests. GridsearchCV helped select parameters for the better scoring models.

**Data**

The database has 253,700 rows of data representing unique BRFSS survey respondents from 2015.

|  |  |
| --- | --- |
| **Column** | **Value Meanings** |
| CholCheck | 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years |
| AnyHealthcare | Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes |
| NoDocbcCost | Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes |
| Education | Education level (EDUCA see codebook) scale 1-6 (1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)) |
| Income | Income scale (INCOME2 see codebook) scale 1-8 (1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more) |
| Fruits | Consume Fruit 1 or more times per day 0 = no 1 = yes |
| Veggies | Consume Vegetables 1 or more times per day 0 = no 1 = yes |
| HighBP | 0 = no high BP 1 = high BP |
| HighChol | 0 = no high cholesterol 1 = high cholesterol |
| PhysActivity | physical activity in past 30 days - not including job 0 = no 1 = yes |
| DiffWalk | Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes |
| BMI | Body Mass Index |
| MentHlth | Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days |
| PhysHlth | Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days |
| GenHlth | Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor |
| Age | 13-level age category (\_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older |
| Smoker | Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes |
| Stroke | (Ever told) you had a stroke. 0 = no 1 = yes |
| HeartDiseaseorAttack | coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes |
| HvyAlcoholConsump | Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes |
| Sex | 0 = female 1 = male |

**Tools**

* Sklearn – classification model building, gridsearchCV, confusion matrix
* Pandas for EDA

**Findings**

* Random forests, extra trees, and decision trees had the best scores for the target metric of recall (63% & 53%).
  + BMI, social background, and age appeared to be the most important risk factors.
  + Surprisingly, alcohol consumption & history of heart disease/stroke appeared to be the least impactful features.
* Feature engineering improved all models. The trees did not prefer the untreated data.

**Future Work**

* Increased feature engineering
* Apply GridSearchCV optimization to more models
  + Linear regression had poor recall scores, but a good ROC\_AUC curve
* Try multinomial instead of 2 split binomial databases

**Communication**

Slides will be available on my [Github](https://github.com/Franchalanche/Flight_Price_Prediction_India).