What are the Risk Factors for Type 2 Diabetes?

Anna Soloveva

About Diabetes

Chronic diseases which can lead to stroke, kidney failure, heart diseases and death

Seventh major cause of death in the US (29.1 Million in 2012 diagnosed with diabetes)

Three main types of diabetes: Type 1, Type 2 (90-95%) and gestational

Cost of estimated diabetes $327 billion in 2017

About Data

Data is downloaded from Behavioral Risk Factor Surveillance System (BRFSS) with 279 variables (464,644 records) for 2014

Dependent Variable is binary classification of Yes or No answer on **"Have you ever been told you have diabetes?"** question

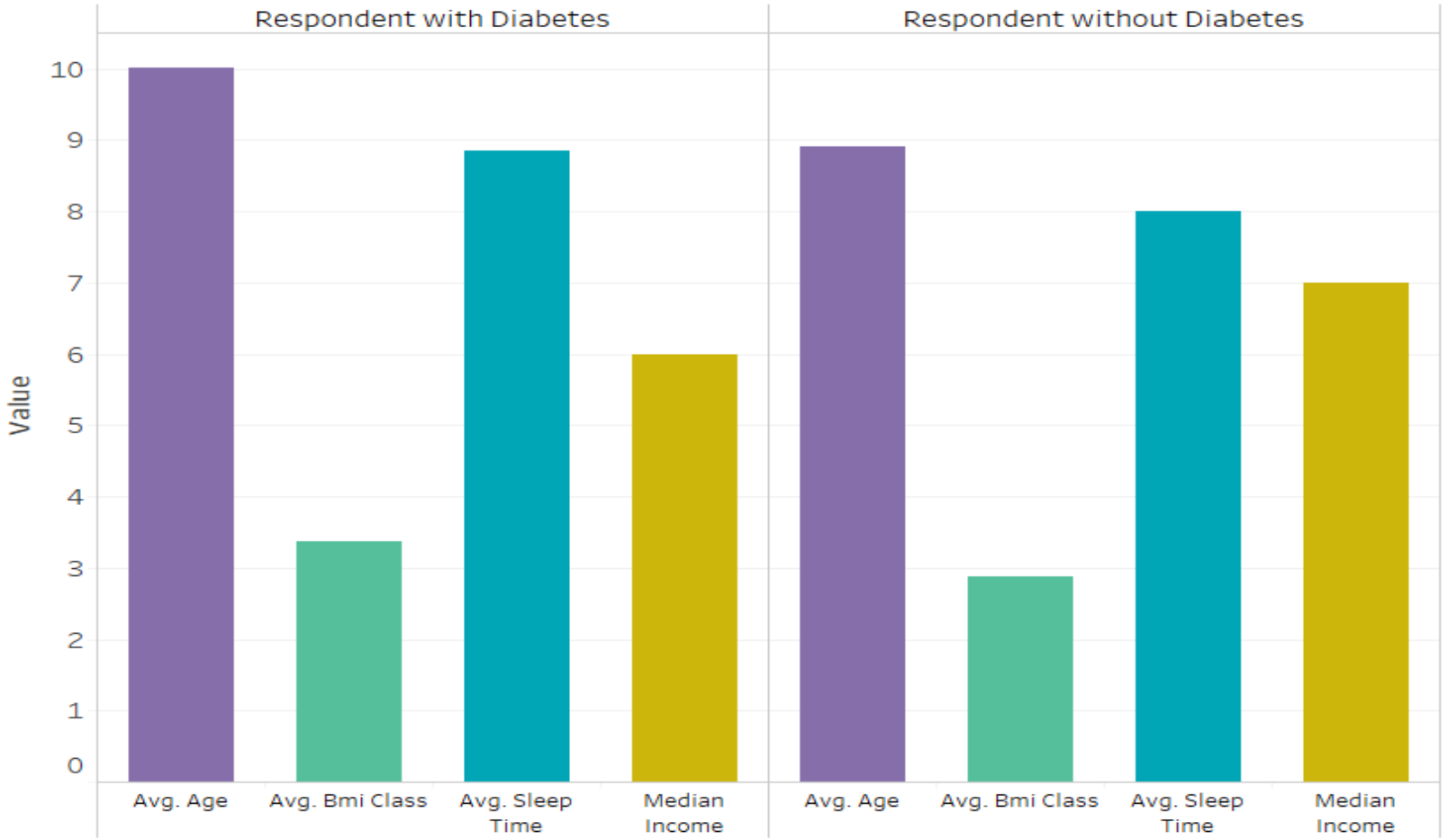
Predictor Variables are 26 personal and general health related characteristics such as General Health, BMI, Age and Sleep Time.

Missing Values have been excluded from analysis: 175,853 records used

Target Variable represents: 85% (No) / 15% (Yes)

Data Understanding

Diebetes is found in patients with higher Age, Body Mass Index and Sleep Time and with lower Income



Random Forest Model

Results:

ROC AUC: **0.78**

Out-of-Bag Error score: **0.84**

Cross-Validation Score: **0.80**

Defined feature columns and independent variable which is *diabetes3*

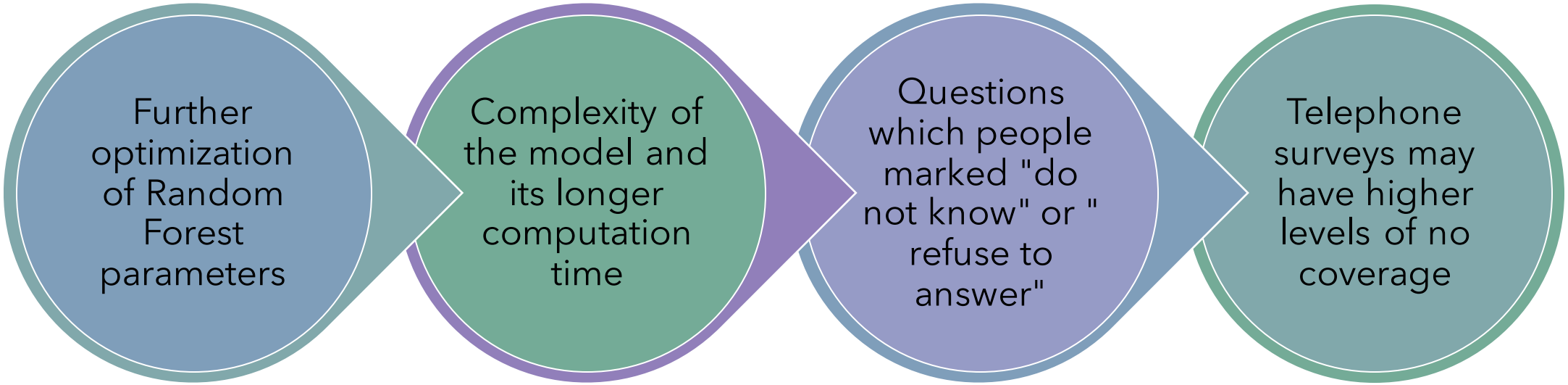
Splitted the data on training and testing sets assigning 70% and 30% respectevely

Used various statistical libraries along with Scikit-Learn library which provides Random Forest classifier function

Fitted the model on the Training Data

Used fitted model to make predictions on Testing Data

Limitations and Challenges



Conclusion

- The model defined important fetures which could be used in early diagnosis and treatment.
- The model can be used to reduce medical costs

Feature	Importance
Income	0.09
Sleep Time	0.09
Age	0.08
General Health	0.07
Metropolitan Status Code	0.06
Education	0.06
Health Care Coverage	0.05
Body Mass Index	0.05
Mental Health	0.05
Marital Status	0.05

References

I. Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev Chronic Dis 2019;16:190109. DOI: http://dx.doi.org/10.5888/pcd16.190109external ic on.

II. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med 2011;9(1):103. https://bmcmedicine.biomedcentral.com/articles/10.1186/1741-7015-9-103

III. Larson W. Insights into health and behavior using data from the CDC. https://github.com/winstonlarson/brfss

IV. Nelson J. Decesion Trees. Adopted from Chapter 8 of An Introdutction to Statistical Learning. http://faculty.marshall.usc.edu/gareth-james/