# What are the Risk Factors for Type 2 Diabetes?

Anna Soloveva

# About Diabetes

Chronic disease 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 year 2017 alone

# About Data

**Original data** has 279 variables and 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

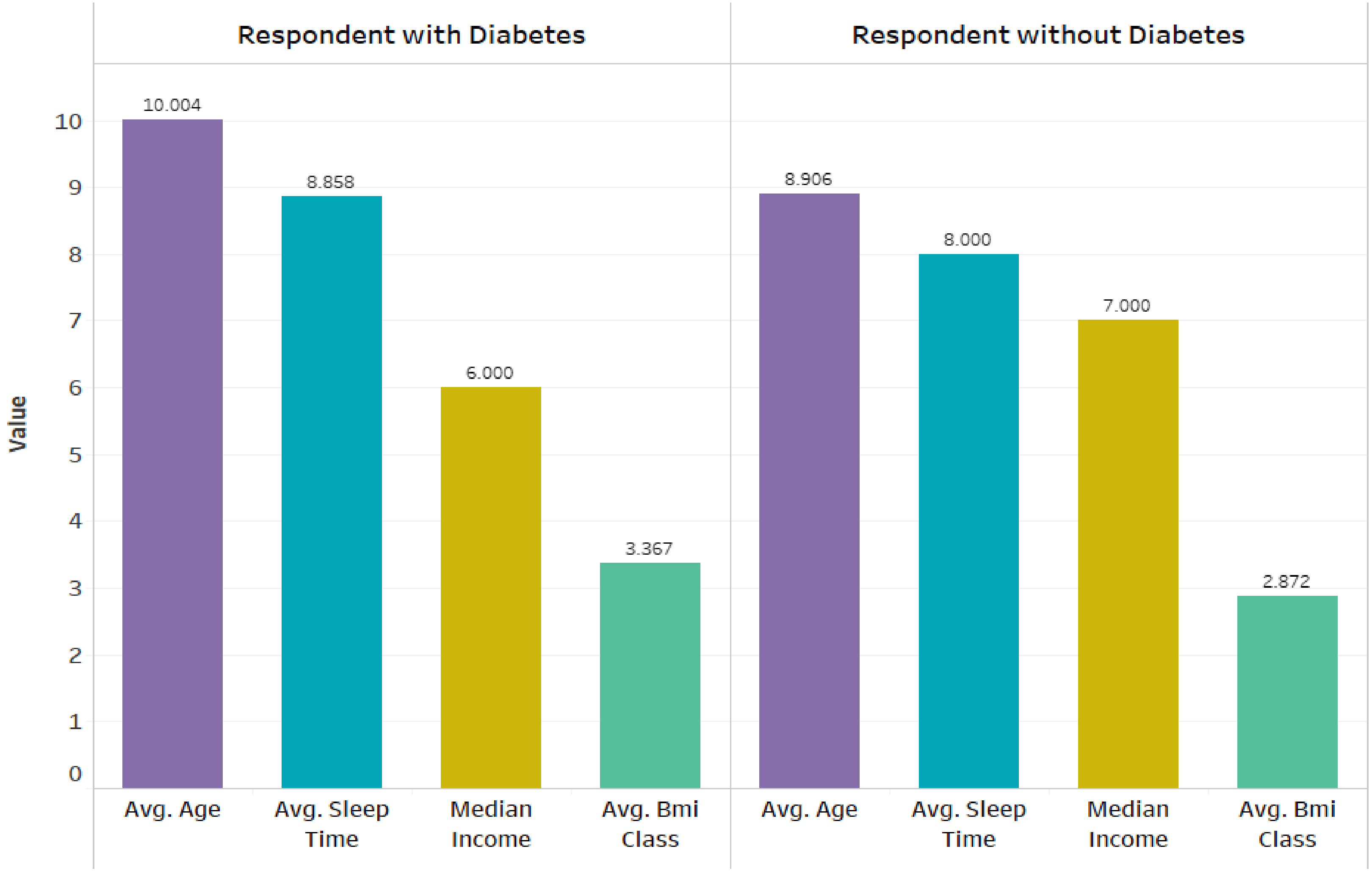
**Independent variables** are 26 personal and general health related characteristics such as General Health, BMI, Age and Sleep Time

**Missing values** have been excluded from the analysis: 175,853 records left

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

# Data Understanding

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



# Random Forest Model

- Defined feature columns and dependent variable
- Separated the data on training and testing sets assigning 70% and 30% respectively
- Used 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
- Cross-Validation Score is **0.83** with 10 k-fold buckets

# Limitations and Challenges

- Further optimization of Random Forest Model parameters
- Complexity of the model and its longer computation time
- Questions which people marked "do not know" or " refuse to answer"
- Telephone surveys may have higher levels of no coverage

# Conclusion

- The model defined important features 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 icon.

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/