9-1-2018

A Data-Driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning

Stacey Miertschin
Winona State University

A Data-driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning

An Dinh, Stacey Miertschin, and Amber Young
Mentor: Somya Mohanty

ASA REU at UNCG

October 7, 2018

Diabetes and Cardiovascular Disease

Diabetes and heart disease are two of the most prevalent chronic diseases that lead to death in the United States.

As of 2015...

- **1 in every 4** deaths are caused by cardiovascular disease
- About **9%** of US has diabetes
  - About **34%** of people in the US population have pre-diabetes
  - Of those with pre-diabetes **90%** were unaware of their condition

Center for Disease Control and Prevention

NHANES Background

National Health and Nutrition Examination Survey

- Nationally representative sample of about 5,000 people each year from 1999-2016
- Includes information such as:
    - Demographics
    - Health-related questionnaire
    - Dietary information
    - Laboratory results
    - Physical examination

Approach of Prior Research

Yu et al. (2010)
Semerdjian and Frank (2017)

- Prior process
  - **Excluded:** Patients under 20 years old and pregnant patients
  - **Data time frame:** The three waves from 1999-2004
  - **Feature selection:** Chose 14-16 features that tend to be associated with diabetes
- Limitations
  - Formal selection of important features
  - Few observations
  - Only predicted diabetes

Our Approach

Our Approach

- Examined more years 1999-2014 $\Rightarrow$ more observations
- Predicted cardiovascular disease as well as diabetes
- Used data-driven feature selection methods

Data Gathering and Cleaning

- Challenges:
  - Many datasets
  - Missing data from conditional questions
  - Discontinuity: variables differ from cycle to cycle
- Data Preprocessing
  - Number of features after preliminary selection is 189 out of approx. 3900 features
  - Excluded subjects under 20, pregnant subjects, and subjects who did not complete an examination

Who is considered diabetic?

Criteria		Classification
If they answered yes to "Have you been told by a doctor that you have diabetes" or plasma glucose $\geq$ 126 mg/dl	$\Rightarrow$	Diabetic
If they answered no, but their plasma glucose $\geq$ 126 mg/dl	$\Rightarrow$	Undiagnosed diabetic
If their plasma glucose was 100–125 mg/dl	$\Rightarrow$	Pre-diabetic
If plasma glucose $\leq$ 100 mg/dl	$\Rightarrow$	Not diabetic

Diabetes Cases Labels

- Case I: Predicting diabetics
- Case II: Predicting undiagnosed diabetics and pre-diabetics

Classification	Diab. Case I	Diab. Case II
Diabetic	1	Excluded
Undiagnosed diabetic	1	1
Pre-diabetic	0	1
Not diabetic	0	0

Who has cardiovascular disease?

Criteria		Classification
If they answered yes to having had one of the following: • Congestive heart failure • Coronary heart disease • Heart attack • Stroke	$\Rightarrow$	Having heart diseases
If they answered no to all of those questions	$\Rightarrow$	Not having heart diseases

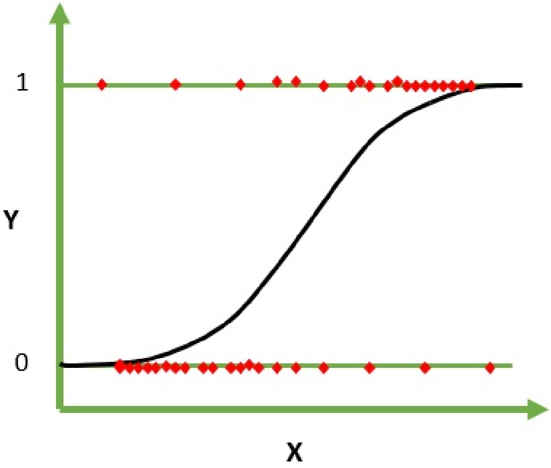
Creating the datasets

Year	Case	Observations	Variables	No. of 1s	No. of 0s
1999-2014	Diab. Case I	21,131	123	15,599	5,532
1999-2014	Diab. Case II	16,426	123	9,944	6,482
2003-2014	Diab. Case I	16,443	168	11,977	4,466
2003-2014	Diab. Case II	12,636	168	7,503	5,133
2007-2014	Cardio	8,459	131	7,012	1,447

Table 1: The table summarizes the structure of the datasets used for diabetes classification.

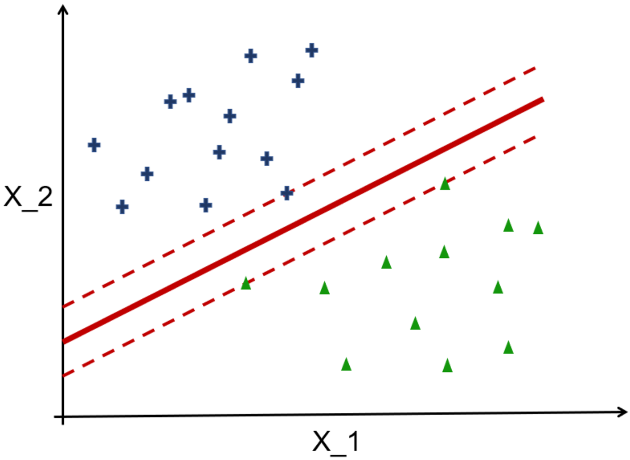
Machine Learning Models

Logistic Regression



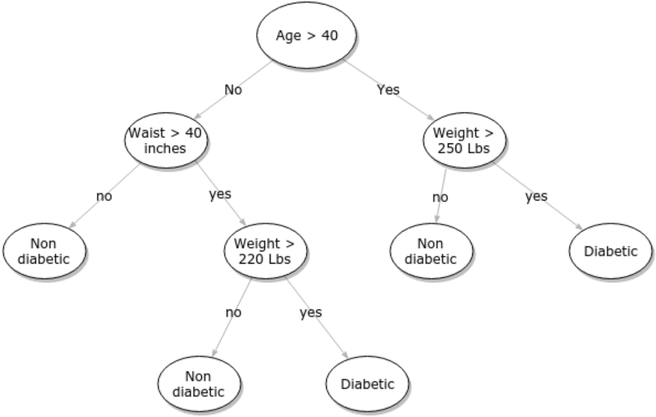
Cox (1958)

Support Vector Machine



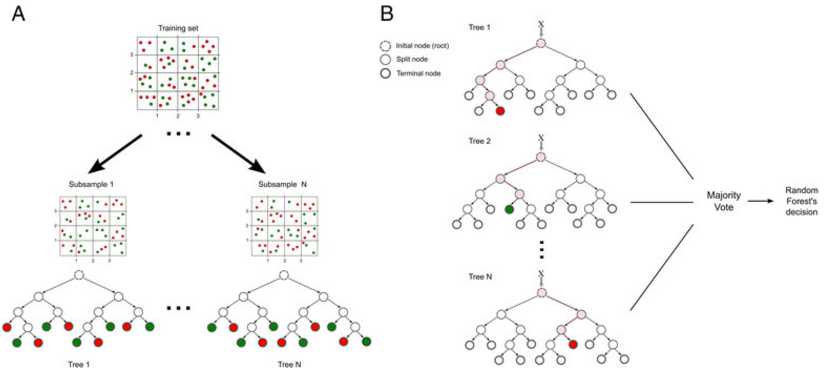
Cortes and Vapnik (1995)

Decision Tree



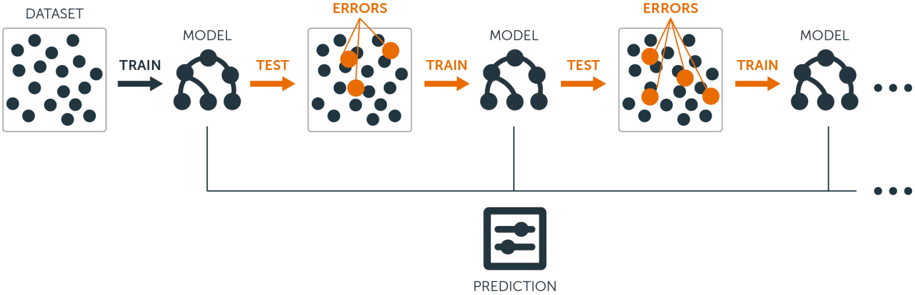
Quinlan (1986)

Random Forests



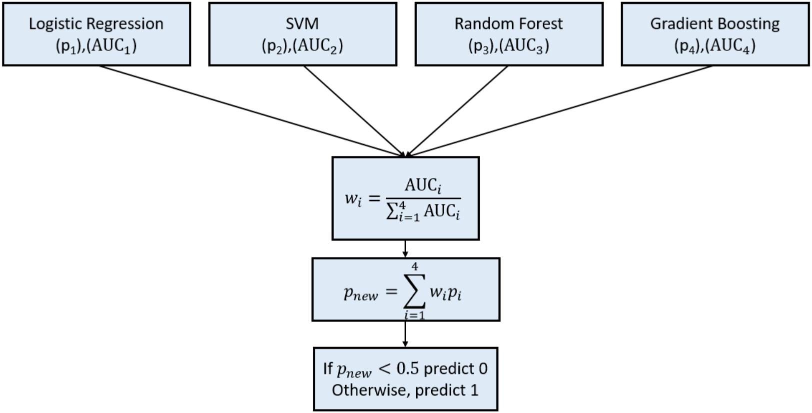
Ho (1995)

Gradient Boosting



Chen and Guestrin (2016)

Ensemble Classifier



Results

Diabetes Model Results without Lab Data

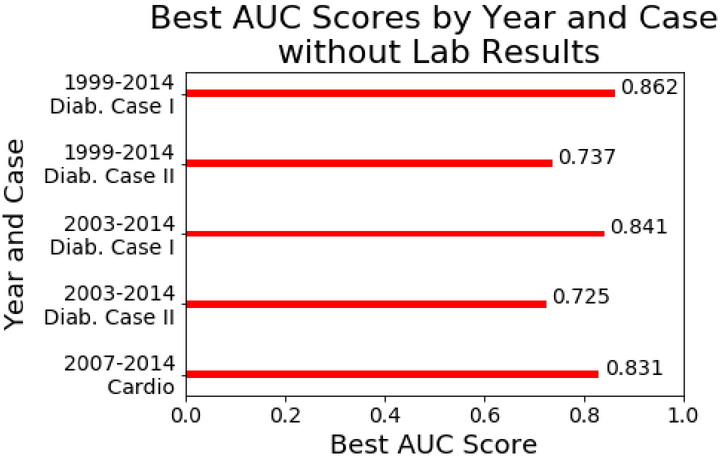


Figure 1: AUC scores of machine learning models without lab results

Diabetes Model Results With Lab Data

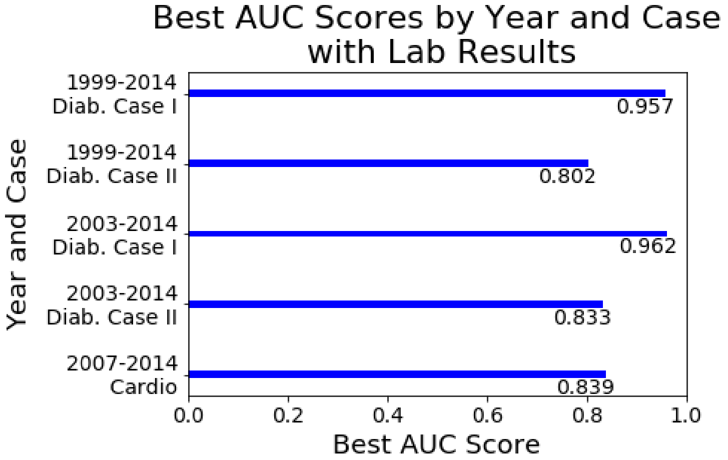


Figure 2: AUC scores of machine learning models with lab results

Diabetes Model Results Comparison

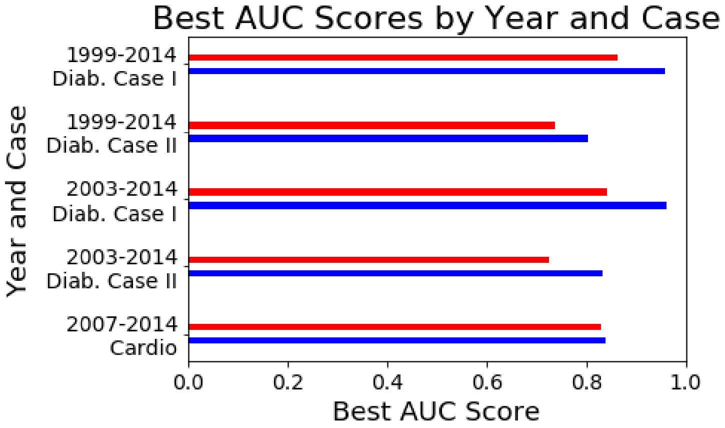


Figure 3: AUC scores of machine learning models with and without lab results

Comparison with Prior Research

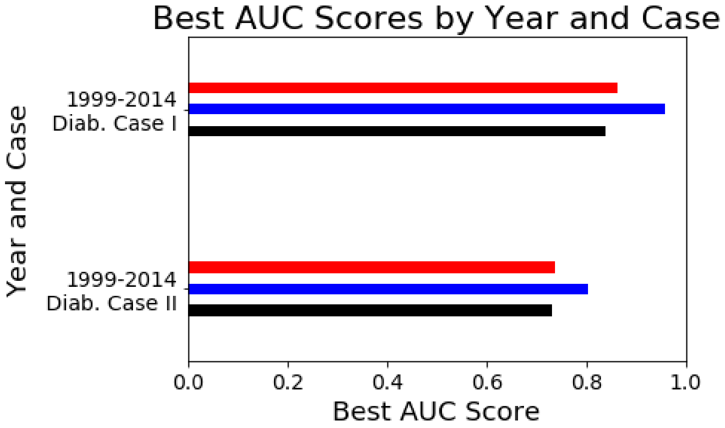


Figure 4: AUC scores of machine learning models with and without lab results compared to prior results

Feature Importance: Diabetes

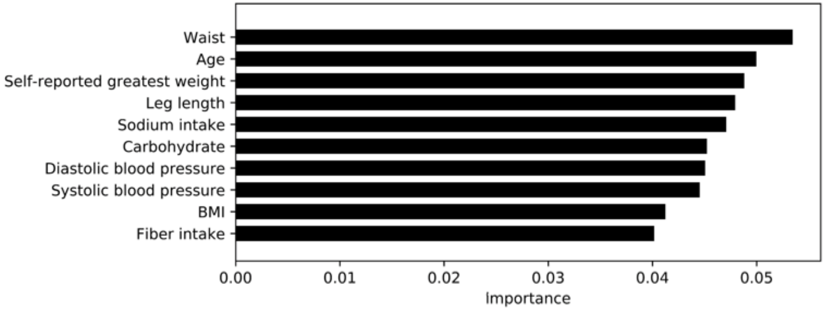


Figure 5: The important features for predicting diabetes

Feature Importance: Cardio

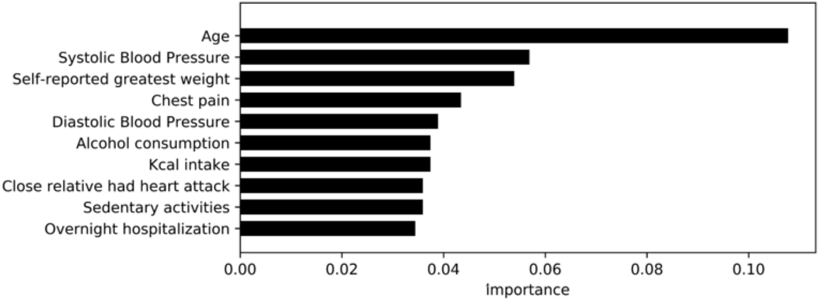


Figure 6: The important features for predicting cardiovascular

Conclusion

- Gradient Boosting and the ensemble classifier are the best performing models
- In predicting diabetes, more observations lead to better predictions
- Lab results greatly improve the models
- Compared to previous papers, our highest AUC score without lab results is about 3% higher
- Future work: Apply models to electronic health records

Acknowledgments

Dr. Somya Mohanty
Dr. Sat Gupta

The University of North Carolina at Greensboro

NSF Grant No. DMS - 1560332

References

(2017). Heart disease fact sheet.

(2017). National diabetes statistics report.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Cortes, C. and Vapnik, V. N. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

Ho, T. K. (1995). Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, volume 1, pages 278–282. IEEE.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Semerdjian, J. and Frank, S. (2017). An Ensemble Classifier for Predicting the Onset of Type II Diabetes. *ArXiv e-prints*.

Yu, W., Liu, T., Valdez, R., Gwinn, M., and Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making*, 10(1):16.