Predicting Diabetes Risk

Fall 2024 - W207 Applied Machine Learning - Final Project

Mayumy Cordova, Hyung Jin Kim, Qiong Zhang, Godsee Joy

UC Berkeley

Introduction: Problem Motivation + Overview

2021 population estimates for U.S. prevalence of diabetes[16]:

15%

Of all U.S. adults had diabetes

23%

Of U.S. adults with diabetes are ***undiagnosed***

38%

Of U.S. adults have prediabetes

Opportunity: help medical practitioners better identify those at risk for diabetes to recommend testing given 23% are undiagnosed using survey questions vs more involved blood tests / diagnostics

Introduction: Problem Motivation + Overview

Previous Literature:

- Risk factors for Type 2 diabetes:
  - Being overweight, higher BMI, obesity[1,8-9]
  - 45+ years old, family history (parent or sibling), physically active less than 3 times a week, being African American/Hispanic/Latino/Native American/some Pacific Islander and Asian American[1]
  - Men are more at risk, with onset at a much lower BMI, but females often have more serious complications[2-3]
  - Smokers are 30-40% more likely to develop Type 2 diabetes than those who don't[4-5]
  - Hypertension is twice as frequent in patients with diabetes compared to those without[6]
  - Diabetes tends to lower good cholesterol and raise bad cholesterol[7]
  - Being uninsured[10]
  - Having lower income, educational attainment, and occupation grade[11-14]
  - Paper referenced doing similar ML prediction task, but for 2015 data: found 74-82% accuracy[15]

Introduction: Problem Motivation + Overview

Key Questions: Can diabetes (Type 2) be predicted with sufficient accuracy based on non-medical, survey data?

- *Specific Aim 1: Whether the accuracy of models trained and tested on 2015 data only can be improved by selecting a proper model or hyperparameter tuning?*
- *Specific Aim 2: Whether the accuracy of models would remain the same if we applied the 2015 trained model on 2023 test data?*
- *Specific Aim 3: Whether the accuracy of models trained by both 2015 and 2023 (combined) data can be improved on 2023 test data by selecting a proper model or hyperparameter tuning?*

Introduction: Problem Motivation + Overview

Analysis Plan:

- Step 1: We process both 2015 data and 2023 data, clean them, conduct exploratory data analysis (EDA) and literature review to select most relevant features
- Step 2: We train multiple models on 2015 data and then apply trained model (i.e., 2015 model) on 2023 data
- Step 3: We combine 2015 and 2023 train, validation, and test data to train same models (i.e., combined model) again and test on combined and 2023 test data

Summary of Results:

- Based on the EDA, there were no significant population changes across variables used between 2015 and 2023.
- We found that most models showed similar performance (70~85% accuracy) on both the 2015 and 2023 data.
- When we applied the 2015 models to the 2023 test data, the accuracy did not change much, but the recall decreased compared to the 2015 test data.
- When we applied the combined models to the 2023 test data, the overall accuracy did not change much and slightly decreased, but the recall improved compared to the results obtained using the 2015 models.
- For the 2023 test data (i.e., more recent data), using models trained on combined data improved the true positive rate (aka helped minimize false negatives). AUC scores were between .61 to .71 (classification threshold of .5) which suggests the model is able to distinguish the two diabetes classes, but there is room for improvement in the modeling process or data.

About the Data + Pre-processing 2015 + 2023

In 1984, the CDC started a state-wide Behavioral Risk Factor Surveillance System (BRFSS), which is an ongoing random digital-dialed telephone survey of noninstitutionalized US adults aged 18+. We used 19 variables to predict diabetes status.

1

Getting the data

- 2015 and 2023 data was pulled from the public CDC BRFSS database as ASCII files
- Specific fields from the survey were pulled based our team's literature review and adaptation of a Kaggle reference notebook and survey coding logic (CC0: Public Domain license)
- 430-440K rows each year

2

Data prep and cleaning

- For both datasets: dropped missing values, recoded numeric variables and categorical variables into smaller categorical or binary features - set all to int64
- For 2023 data, recoded variables to match 2015 data structure
- Left with about 260K rows for each year

3

EDA + Data Splits

- Conducted EDA for both 2015 and 2023 data
- Split 2015 and 2023 datasets into train, validation, and test datasets using a 60%-20%-20% split
- Created combined train, validation, and test sets by joining across both years

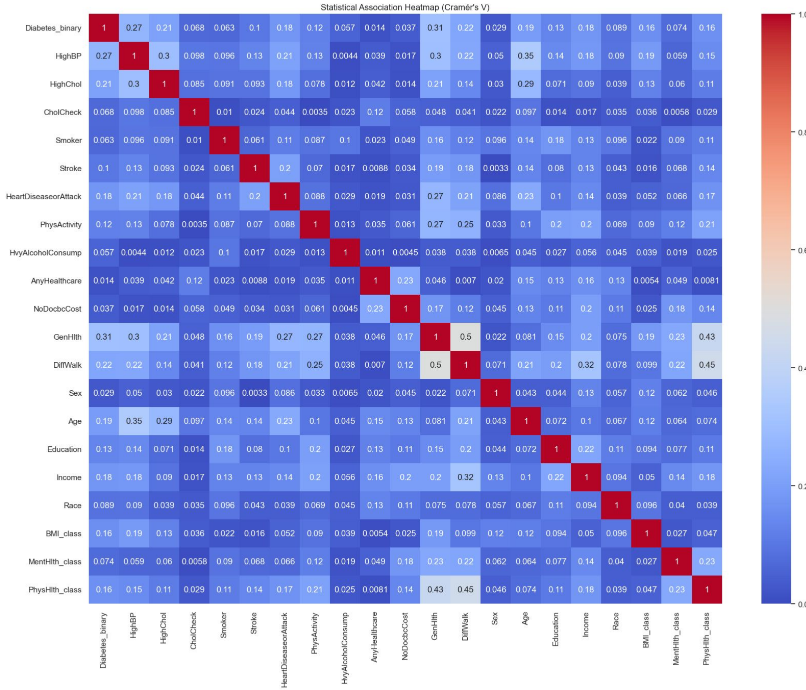
4

Final Prep for Modeling

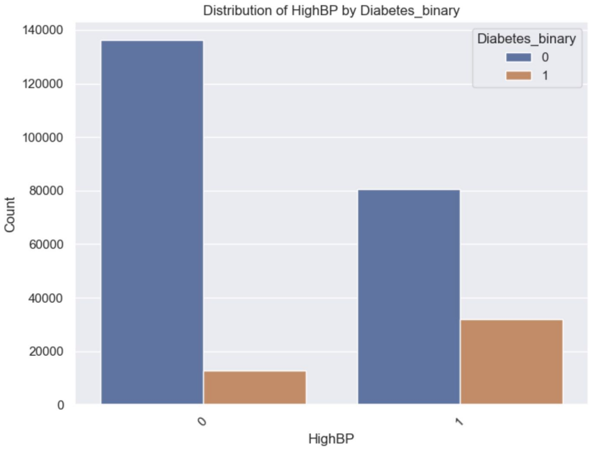
- Target binary variable of diabetes was very unbalanced, only about 17% with diabetes
- Used SMOTE on both 2015 and combined train sets which over-samples the minority class by generating synthetic data

Exploratory Data Analysis: 2015 and 2023

- **High similarity between 2015 and 2023 data**
- **Total Data:**
    - **2015:** 0-Non-diabetic(220756), 1-Diabetic(41662)
    - **2023:** 0-Non-diabetic(216724), 1-Diabetic(44821)
- **Health Conditions** such as high blood pressure, high cholesterol, and physical inactivity are more common in diabetics
- We used bar plots and a correlation heatmap (Cremer's V) to understand the pairwise relationships between all variables

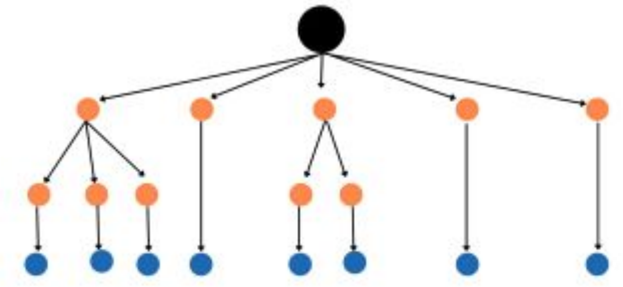

Statistical Association Heatmap (Cramér's V)

	Diabetes_binary	HighBP	HighChol	CholCheck	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	DiffWalk	Sex	Age	Education	Income	Race	BMI_class	MentHlth_class	PhysHlth_class
Diabetes_binary	1	0.27	0.21	0.068	0.063	0.1	0.18	0.12	0.057	0.014	0.037	0.31	0.22	0.029	0.19	0.13	0.18	0.089	0.16	0.074	0.16
HighBP	0.27	1	0.3	0.098	0.096	0.13	0.21	0.13	0.0044	0.039	0.017	0.3	0.22	0.05	0.35	0.14	0.18	0.09	0.19	0.059	0.15
HighChol	0.21	0.3	1	0.085	0.091	0.093	0.18	0.078	0.012	0.042	0.014	0.21	0.14	0.03	0.29	0.071	0.09	0.039	0.13	0.06	0.11
CholCheck	0.068	0.098	0.085	1	0.01	0.024	0.044	0.0035	0.023	0.12	0.058	0.048	0.041	0.022	0.097	0.014	0.017	0.035	0.036	0.0058	0.029
Smoker	0.063	0.096	0.091	0.01	1	0.061	0.11	0.087	0.1	0.023	0.049	0.16	0.12	0.096	0.14	0.18	0.13	0.096	0.022	0.09	0.11
Stroke	0.1	0.13	0.093	0.024	0.061	1	0.2	0.07	0.017	0.0088	0.034	0.19	0.18	0.0033	0.14	0.08	0.13	0.043	0.016	0.068	0.14
HeartDiseaseorAttack	0.18	0.21	0.18	0.044	0.11	0.2	1	0.088	0.029	0.019	0.031	0.27	0.21	0.086	0.23	0.1	0.14	0.039	0.052	0.066	0.17
PhysActivity	0.12	0.13	0.078	0.0035	0.087	0.07	0.088	1	0.013	0.035	0.061	0.27	0.25	0.033	0.1	0.2	0.2	0.069	0.09	0.12	0.21
HvyAlcoholConsump	0.057	0.0044	0.012	0.023	0.1	0.017	0.029	0.013	1	0.011	0.0045	0.038	0.038	0.0065	0.045	0.027	0.056	0.045	0.039	0.019	0.025
AnyHealthcare	0.014	0.039	0.042	0.12	0.023	0.0088	0.019	0.035	0.011	1	0.23	0.046	0.007	0.02	0.15	0.13	0.16	0.13	0.0054	0.049	0.0081
NoDocbcCost	0.037	0.017	0.014	0.058	0.049	0.034	0.031	0.061	0.0045	0.23	1	0.17	0.12	0.045	0.13	0.11	0.2	0.11	0.025	0.18	0.14
GenHlth	0.31	0.3	0.21	0.048	0.16	0.19	0.27	0.27	0.038	0.046	0.17	1	0.5	0.022	0.081	0.15	0.2	0.075	0.19	0.23	0.43
DiffWalk	0.22	0.22	0.14	0.041	0.12	0.18	0.21	0.25	0.038	0.007	0.12	0.5	1	0.071	0.21	0.2	0.32	0.078	0.099	0.22	0.45
Sex	0.029	0.05	0.03	0.022	0.096	0.0033	0.086	0.033	0.0065	0.02	0.045	0.022	0.071	1	0.043	0.044	0.13	0.057	0.12	0.062	0.046
Age	0.19	0.35	0.29	0.097	0.14	0.14	0.23	0.1	0.045	0.15	0.13	0.081	0.21	0.043	1	0.072	0.1	0.067	0.12	0.064	0.074
Education	0.13	0.14	0.071	0.014	0.18	0.08	0.1	0.2	0.027	0.13	0.11	0.15	0.2	0.044	0.072	1	0.22	0.11	0.094	0.077	0.11
Income	0.18	0.18	0.09	0.017	0.13	0.13	0.14	0.2	0.056	0.16	0.2	0.2	0.32	0.13	0.1	0.22	1	0.094	0.05	0.14	0.18
Race	0.089	0.09	0.039	0.035	0.096	0.043	0.039	0.069	0.045	0.13	0.11	0.075	0.078	0.057	0.067	0.11	0.094	1	0.096	0.04	0.039
BMI_class	0.16	0.19	0.13	0.036	0.022	0.016	0.052	0.09	0.039	0.0054	0.025	0.19	0.099	0.12	0.12	0.094	0.05	0.096	1	0.027	0.047
MentHlth_class	0.074	0.059	0.06	0.0058	0.09	0.068	0.066	0.12	0.019	0.049	0.18	0.23	0.22	0.062	0.064	0.077	0.14	0.04	0.027	1	0.23
PhysHlth_class	0.16	0.15	0.11	0.029	0.11	0.14	0.17	0.21	0.025	0.0081	0.14	0.43	0.45	0.046	0.074	0.11	0.18	0.039	0.047	0.23	1


Distribution of HighBP by Diabetes_binary

Models Explored

- Baseline Model
- Model 1: Logistic Regression Model
- Model 2: Logistic Regression Model in TensorFlow
- Model 3: Support Vector Machine
    - *ended up not using after SMOTE*
- Model 4: Decision Tree
- Model 5: Random Forest
- Model 6: Gaussian Naïve Bayes
- Model 7: Neural Network (MLPClassifier Scikit Learn)
- Model 8: Neural Network (TensorFlow)

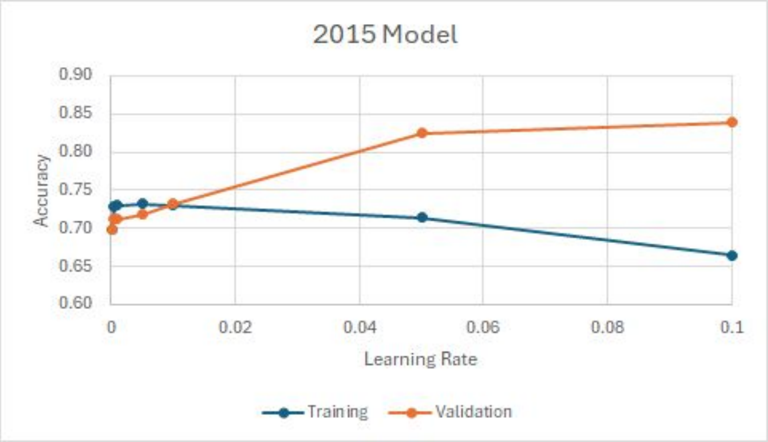


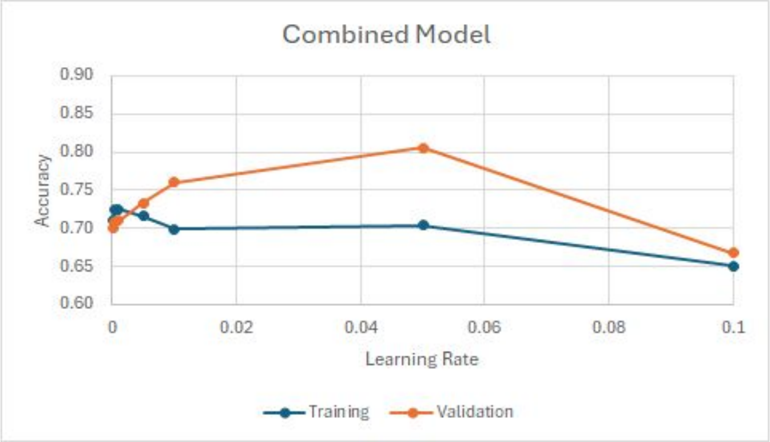
Experiments and Results

- Models with Hyperparameter (HP) Tuning Experiments
    - Logistic Regression (LR) Model in TensorFlow (TF)
    - Decision Tree
    - Random Forest
    - Neural Network (in Scikit Learn and TensorFlow)

Hyperparameter Tuning for Logistic Regression Model in TF

- Grid Search: [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1]
- Objective: Accuracy on Validation Set
- Callback Option: Loss on Validation Set
- Best Learning Rates
  - 2015 Model: 0.1
  - Combined Model: 0.05





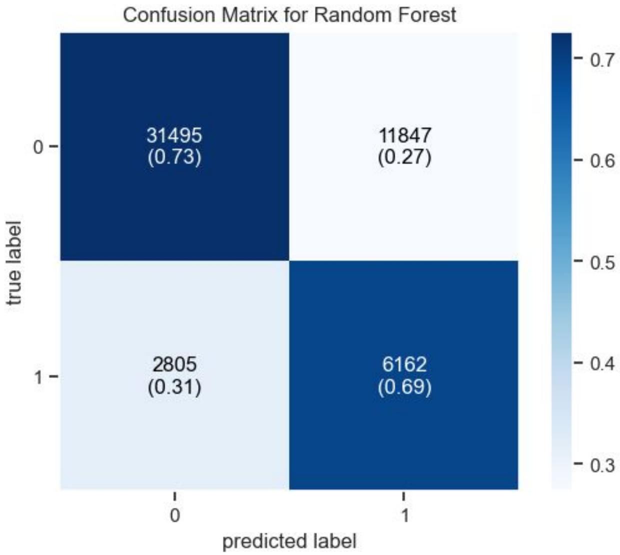
Decision Tree and Random Forest Tuning

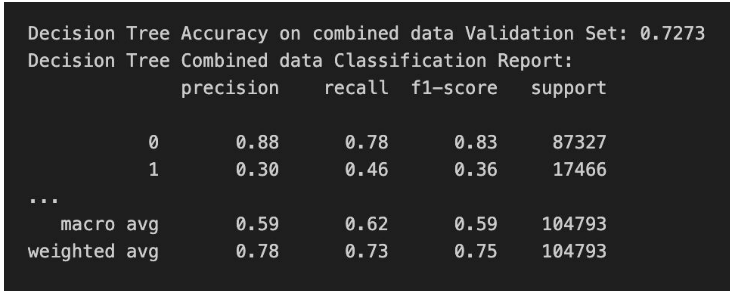
Decision Tree

- Grid search:
  - Criterion: Gini, Entropy
  - Max Depth: 3, 5, 10, none
  - Min Samples Split: 2, 5, 10
  - Min Samples Leaf: 1, 2, 5, 10
  - Max Features: None, sqrt, log2
- Decision Tree for 2015 and Combined best parameters:
  - Criterion: Entropy, Max Depth: None, Max Features: None, Min Samples Leaf: 1, Min Samples Split 2

Random Forest

- Grid Search:
  - n_estimators: [100, 200, 300],
  - max_depth: [None, 10, 20, 30],
  - min_samples_split: [2, 5, 10],
  - min_samples_leaf: [1, 2, 4],
  - bootstrap: [True, False]
- Random Forest for 2015 Model: n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_depth': None, 'bootstrap': False
- Random Forest for Combined Model: 'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 20, 'bootstrap': True



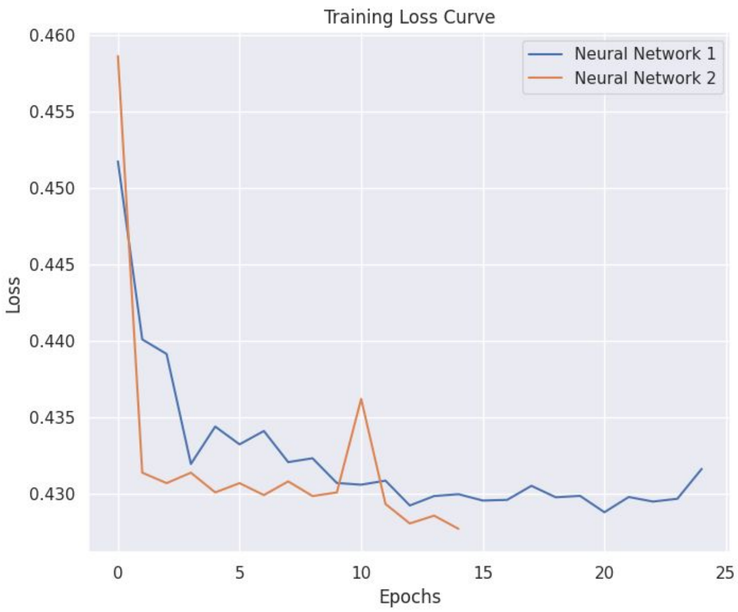


```
Decision Tree Accuracy on combined data Validation Set: 0.7273
Decision Tree Combined data Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.78      0.83     87327
           1       0.30      0.46      0.36     17466
...
   macro avg       0.59      0.62      0.59    104793
weighted avg       0.78      0.73      0.75    104793
```

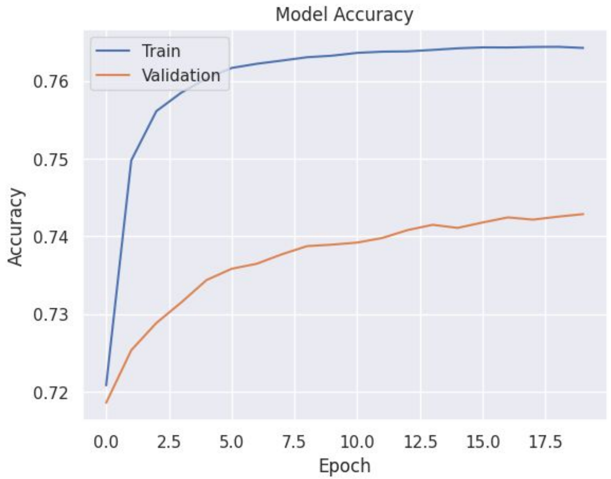
Feedforward Neural Network Tuning (MLPClassifier Scikit Learn)

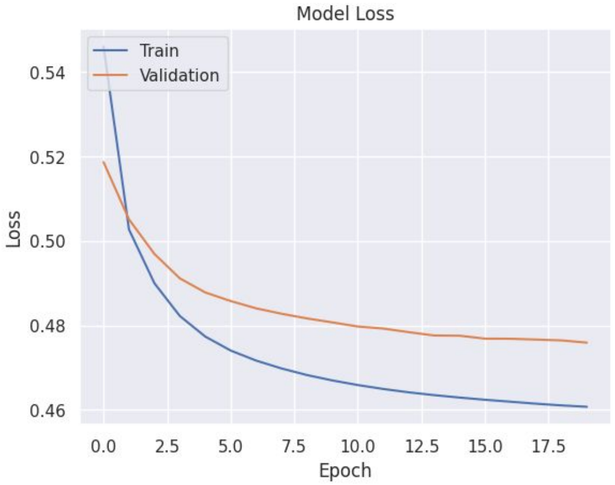
- Manual tuning due to grid search long run times (2015 training loss curves for 2 models on right)
- Optimizing recall slightly over overall accuracy, but second best model had slightly higher accuracy
- Parameter ranges (all relu activation):
  - Hidden Layer Sizes: [(50,), (100,), (100, 50)],
  - Solver: ['adam', 'sgd'],
  - Alpha: [0.0001, 0.001, 0.01, 0.1],
  - Learning Rate: [0.001, 0.01, 0.1, .05, .08]
- Best parameters:
  - Hidden layers: (50,)
  - Solver: adam
  - Alpha: .01
  - Learning rate: .05



Feedforward Neural Network Tuning (TF)

- Manual tuning due to grid search long run times (2015 model accuracy and loss curves on right)
- Optimizing recall slightly over overall accuracy
- Parameter ranges:
    - Learning rates: .001, .01, .005, .0005
    - Units: 5, 8, 10, 16
    - Hidden layers: 1, 2
- Best parameters:
    - Learning rate: .0005
    - Optimizer: adam
    - Activator: relu
    - Units per hidden layer: 5
    - Hidden Layers: 1





MODELS SUMMARY

	2015 MODEL				COMBINED MODEL			
MODEL NAME	2015 TEST		2023 TEST		COMBINED TEST		2023 TEST	
	ACCURACY	RECALL	ACCURACY	RECALL	ACCURACY	RECALL	ACCURACY	RECALL
Baseline - No Diabetes	0.8420	-	0.8290	-	0.8350	-	0.8290	-
Logistic Regression	0.7276	**0.7433**	0.7301	**0.7071**	0.7202	**0.7386**	0.7223	**0.7215**
Logistic Regression in TensorFLow (HP)	**0.8437***	0.0596	**0.8301***	0.0306	**0.8067***	0.4385	**0.8101***	0.4026
Decision Tree (HP)	0.7414	0.4361	0.7228	0.4160	0.7273	0.4670	0.7296	0.4630
Random Forest	0.7652	0.4607	0.7590	0.4227	0.7215	0.7091	0.7232	0.6938
Gaussian Naive Bayes Classifiers	0.6607	**0.8042***	0.6762	**0.7624***	0.6551	**0.7954***	0.6637	**0.7786***
Neural Network (HP)	0.7435	0.6506	**0.7404**	0.6310	**0.7404**	0.6437	**0.7392**	0.6334
Neural Network in TensorFLow (HP)	**0.7496**	0.6503	0.7303	0.6503	0.7284	0.6942	0.7226	0.6902

Conclusions

- An exercise in model drift: It is important to consider human behavior changes that may affect population data over time, and recognize that models will need to be closely monitored and updated to maintain performance results
- We had hypothesized that combining 2015 and 2023 data to train a new model would greatly improve accuracy on the latest, 2023 data
- However, overall accuracy slightly decreased while recall (sensitivity) increased
- We focus on recall in our results in addition to overall accuracy because:
    - False negatives are more harmful for an individual than a false positive (i.e., having diabetes but given a result that one does not)
    - Treatment for Type 2 diabetes would not begin until further blood work was done
    - Initial recommendations are often around lifestyle and dietary changes
    - Precision scores though were roughly in the 30% range (not shown in slides but in codebook) - which can definitely be improved even though we prioritized recall and overall accuracy
- We were also surprised that model performances were so similar
    - Linear models (logistic regression and Gaussian Naive-Bayes) were top performers and neural networks
    - Logistic Regression (Scikit Learn) was the most consistent performer
- Overall, predicting diabetes status with non-medical survey data seems like a promising area that could aid social workers and medical practitioners in helping close the “undiagnosed” gap of adults living with Type 2 diabetes

Limitations and Areas for Future Study

- **Survey data:** response bias likely, unknown how truthful people are in responses + memory recall which limits data quality
- **Subgroup analysis:** Survey data is mostly White respondents, and survey weights / stratification were not used, but if implemented could lead to more US adult generalizable results. Importantly, the model may not perform equally on all groups. Subgroup analysis by gender, age, and race at a minimum should be done if such a model were to ever be used in clinical / social settings.
- **Missing variables:** Dataset could be expanded to collect family and medical history, and diet information; adding geographic data could help inform more targeted outreach
- **Feature importance analysis**: SHAP analysis to understand which features are carrying most predictive importance, double check this aligns with literature
- **Slight deviations in repeatability:** While we include in the code the random seeds selected for all models and experiments, we note there are slight changes that occur each run due to the initialization process and during SMOTE - we observed small changes from less than 1 to about a couple percentage point differences in evaluation metrics run-to-run
- **Exploring classification thresholds:** explore values other than .5, see how AUC scores, precision, recall, and overall accuracy change
- **No global data:** Data is from US respondents, cannot generalize outside the US
- **Data processing improvements:**
  - Include binary year indicator in combined dataset to give model more information to distinguish between population differences between 2015 and 2023 samples - perhaps combining without including info that they came from 2 different populations could help
  - Consider different training balancing techniques: SMOTE created synthetic data to oversample the minority class while others like NearMiss under-samples the majority class, but other techniques should be explored (e.g., SMOTE resulted in a 50-50 split of diabetic and non-diabetics in train and the interpolation may not have been the best approach)
- **Avoided certain demographic features:** We intentionally did not use certain demographic features like race in our model unlike the research paper referenced did as there are ethical concerns on using these traits to predict disease risk. Race is serving as a proxy for other factors affecting lifestyle, diet for ethnic groups experiencing diabetes at higher rates.

Thank You!

References

1. Centers for Disease Control and Prevention. (2024, May 15). *Diabetes Risk Factors*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/risk-factors/index.html
2. Centers for Disease Control and Prevention. (2024, May 15). *Diabetes and Men*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/risk-factors/diabetes-and-men.html
3. Simmons, H. (2022, October 11). *Diabetes in Men versus Women*. News Medical Life Sciences. https://www.news-medical.net/health/Diabetes-in-Men-versus-Women.aspx
4. Centers for Disease Control and Prevention: National Center for Chronic Disease Prevention and Health Promotion (US) - Office on Smoking and Health (US). (2010). *2010 Surgeon General's Report: How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease*. Centers for Disease Control and Prevention. https://archive.cdc.gov/#/details?url=https://www.cdc.gov/tobacco/sgr/2010/index.htm
5. U.S. National Library of Medicine NIH. (2014). *The Health Consequences of Smoking—50 Years of Progress A Report of the Surgeon General*. National Center for Biotechnology Information (NCBI). https://www.ncbi.nlm.nih.gov/books/NBK179276/pdf/Bookshelf_NBK179276.pdf#page=592
6. Petrie, J. R., Guzik, T. J., & Touyz, R. M. (2018). Diabetes, Hypertension, and Cardiovascular Disease: Clinical Insights and Vascular Mechanisms. *The Canadian journal of cardiology*, *34*(5), 575–584. https://doi.org/10.1016/j.cjca.2017.12.005
7. *Cholesterol and Diabetes*. American Heart Association. (2024, April 2). https://www.heart.org/en/health-topics/diabetes/diabetes-complications-and-risks/cholesterol-abnormalities--diabetes
8. Medhi, G. K., Dutta, G., Borah, P., Lyngdoh, M., & Sarma, A. (2021, January 17). *Prevalence of diabetes and its relationship with body mass index among elderly people in a rural area of Northeastern State of India*. Cureus. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7886600/
9. *Understanding excess weight and its role in type 2 diabetes*. Honor Health. (2011). https://www.honorhealth.com/medical-services/bariatric-weight-loss-surgery/patient-education-and-support/comorbidities-type-2-diabetes
10. Casagrande SS, Park J, Herman WH, et al. Health Insurance and Diabetes. 2023 Dec 20. In: Lawrence JM, Casagrande SS, Herman WH, et al., editors. Diabetes in America [Internet]. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK); 2023-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK597725/
11. Saydah, S., & Lochner, K. (2010). Socioeconomic status and risk of diabetes-related mortality in the U.S. *Public health reports (Washington, D.C. : 1974)*, *125*(3), 377–388. https://doi.org/10.1177/003335491012500306
12. National Center for Chronic Disease Prevention and Health Promotion (U.S.). Division of Diabetes Translation. (2018). Diabetes report card 2017.
13. Borrell, L. N., Dallo, F. J., & White, K. (2006). Education and diabetes in a racially and ethnically diverse population. *American journal of public health*, *96*(9), 1637–1642. https://doi.org/10.2105/AJPH.2005.072884
14. Hill-Briggs, F., Adler, N. E., Berkowitz, S. A., Chin, M. H., Gary-Webb, T. L., Navas-Acien, A., Thornton, P. L., & Haire-Joshu, D. (2020). Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes care*, *44*(1), 258–279. Advance online publication. https://doi.org/10.2337/dci20-0053
15. 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/pcd16.190109
16. Centers for Disease Control and Prevention. (2024c, May 15). National Diabetes Statistics Report. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/php/data-research/index.html

NeurIPS Checklist

1. For all authors...
   - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes, checked the intro and conclusion again after finalizing all results**
   - (b) Have you read the ethics review guidelines and ensured that your paper conforms to them? (Please read the ethics review guidelines) **Yes, code is shared in Github and heavily commented, links to sources and references provided, data was publicly available and no human subjects were used in this research**
   - (c) Did you discuss any potential negative societal impacts of your work? **Yes, discussed briefly in terms of making sure subgroup analysis is done if any model like this was to be deployed in the real world**
     - For more information, see this unofficial guidance and other resources at the broader impacts workshop at NeurIPS 2020.
   - (d) Did you describe the limitations of your work? **Yes, one slide dedicated and some are commented in the code book directly**
2. If you ran experiments...
   - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **Yes, in codebook**
     - Please see the NeurIPS code and data submission guidelines for more details.
   - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes, in codebook**
   - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **Yes, in limitations and seeds are noted in code**
   - (d) Did you include the amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **Primarily used Google Colab free version, models all ran with CPU accelerator**
3. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   - (a) If your work uses existing assets, did you cite the creators? **Yes, in data slide and in code notebook**
   - (b) Did you mention the license of the assets? **Yes, in data slide**
   - (c) Did you include any new assets either in the supplemental material or as a URL? **Yes, we document clearly in code book all steps needed to transform the data into the way we used for our models**
   - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **N/A, leveraged public CDC dataset where consent was obtained in their processes**
   - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes, no PII or offensive content**

Team Contributions + GitHub Repo

Areas	Mayumy	Godsee	Qiong	Hyung Jin
Prior Research & Discovery	●	●	●	●
Data Wrangling	●	●	●	●
Exploratory Data Analysis	●	●	●	●
Model & Data Selection Experiments	●	●	●	●
Project Management & General Strategy	●	●	●	●
Primary Notebook Assembly	●	●	●	●
GitHub & Final Presentation Assembly	●	●	●	●

Github Repo:
https://github.com/godsee-j/mids_207_diabetes_prediction.git