Y

Diabetes Prediction

YueFeng Xue & Irina Lee

Berkeley
UNIVERSITY OF CALIFORNIA

Motivation

How can machine learning be used to help predict whether people have diabetes or prediabetes?

Symptoms being mild and developing over time

Take Years to diagnostic

#8 Ranked Cause of Death

Y

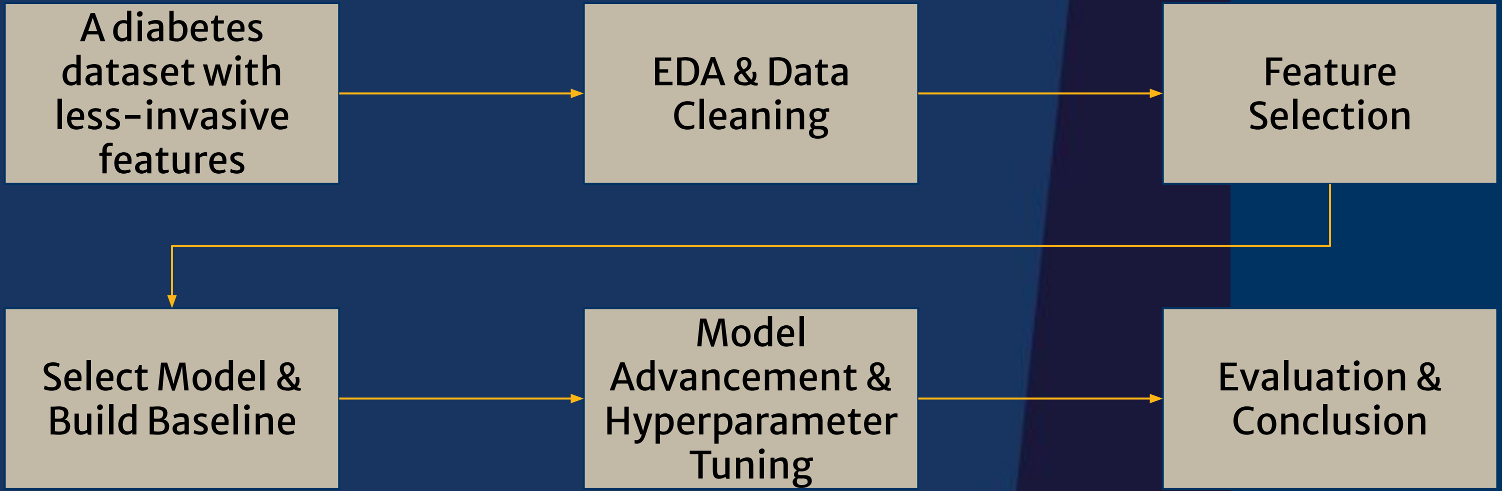
Research Conducted in this Area

Promising Results in Diabetes Prediction Using Machine Learning

Increasing Interests in Predicting Using Non-Invasive Data

Looking for Other Cost & Time Efficient Approach

Our Plan



A diabetes dataset with less-invasive features

EDA & Data Cleaning

Feature Selection

Select Model & Build Baseline

Model Advancement & Hyperparameter Tuning

Evaluation & Conclusion

Y

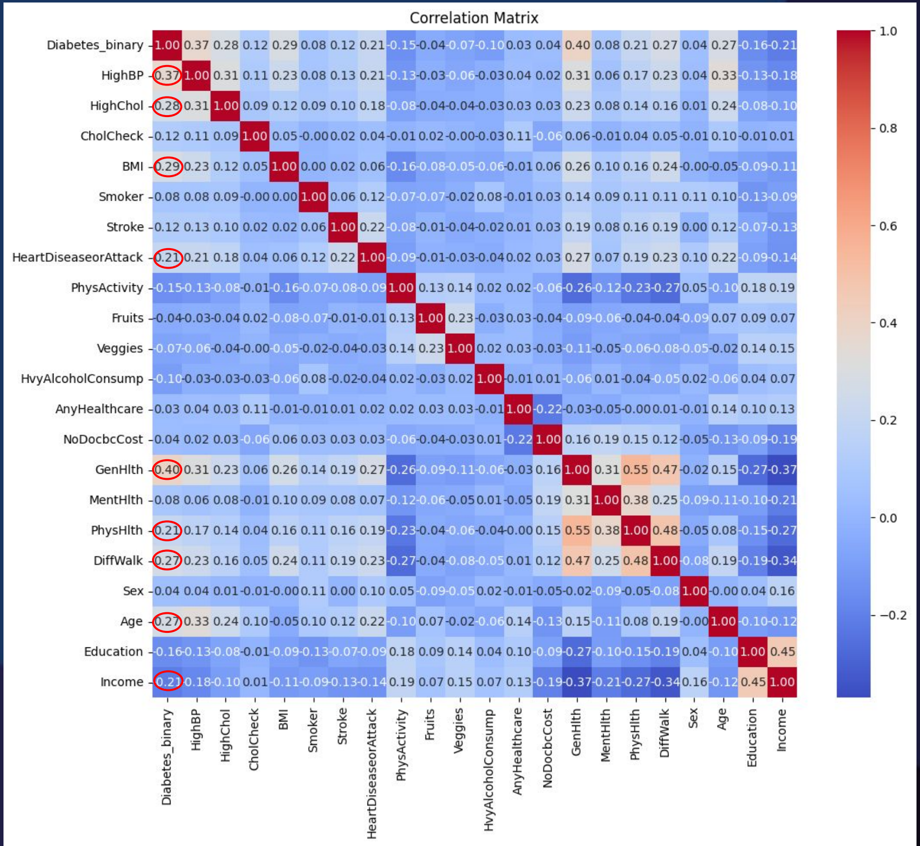
Data

Category	Description
Method of Data Collection	Health-related telephone survey that is collected annually by the CDC in 2015
Number of Features	22
Outcome Feature	Diabetes_binary (0: no diabetes; 1: prediabetes or diabetes)
Outcome Feature Balance	Yes, 50/50 Split
Missing Values	0
Total Rows	70692
Duplicated Rows	1635

Berkeley
UNIVERSITY OF CALIFORNIA

Y

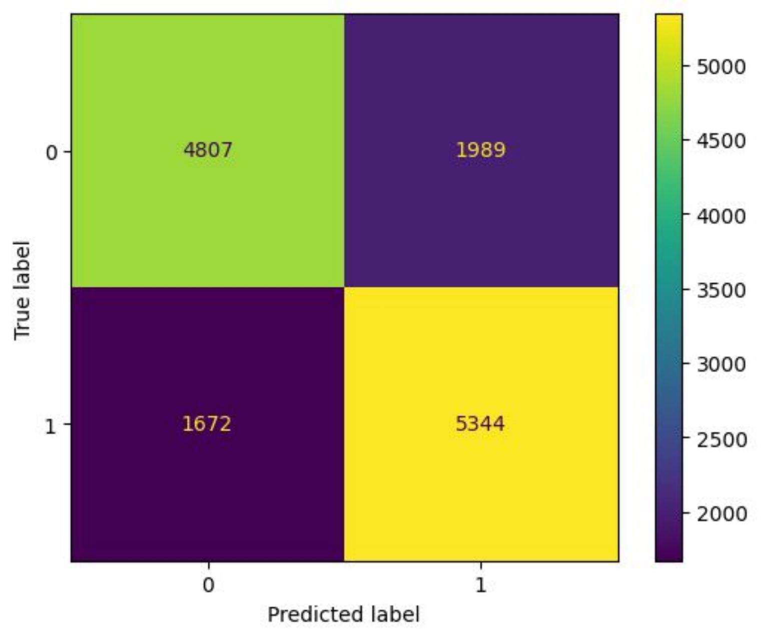
Correlation Matrix



Correlation Matrix

	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
Diabetes_binary	1.00	0.37	0.28	0.12	0.29	0.08	0.12	0.21	-0.15	-0.04	-0.07	-0.10	0.03	0.04	0.40	0.08	0.21	0.27	0.04	0.27	-0.16	-0.21
HighBP	0.37	1.00	0.31	0.11	0.23	0.08	0.13	0.21	-0.13	-0.03	-0.06	-0.03	0.04	0.02	0.31	0.06	0.17	0.23	0.04	0.33	-0.13	-0.18
HighChol	0.28	0.31	1.00	0.09	0.12	0.09	0.10	0.18	-0.08	-0.04	-0.04	-0.03	0.03	0.03	0.23	0.08	0.14	0.16	0.01	0.24	-0.08	-0.10
CholCheck	0.12	0.11	0.09	1.00	0.05	-0.00	0.02	0.04	-0.01	0.02	-0.00	-0.03	0.11	-0.06	0.06	-0.01	0.04	0.05	-0.01	0.10	-0.01	0.01
BMI	0.29	0.23	0.12	0.05	1.00	0.00	0.02	0.06	-0.16	-0.08	-0.05	-0.06	-0.01	0.06	0.26	0.10	0.16	0.24	-0.00	-0.05	-0.09	-0.11
Smoker	0.08	0.08	0.09	-0.00	0.00	1.00	0.06	0.12	-0.07	-0.07	-0.02	0.08	-0.01	0.03	0.14	0.09	0.11	0.11	0.11	0.10	-0.13	-0.09
Stroke	0.12	0.13	0.10	0.02	0.02	0.06	1.00	0.22	-0.08	-0.01	-0.04	-0.02	0.01	0.03	0.19	0.08	0.16	0.19	0.00	0.12	-0.07	-0.13
HeartDiseaseorAttack	0.21	0.21	0.18	0.04	0.06	0.12	0.22	1.00	-0.09	-0.01	-0.03	-0.04	0.02	0.03	0.27	0.07	0.19	0.23	0.10	0.22	-0.09	-0.14
PhysActivity	-0.15	-0.13	-0.08	-0.01	-0.16	-0.07	-0.08	-0.09	1.00	0.13	0.14	0.02	0.02	-0.06	-0.26	-0.12	-0.23	-0.27	0.05	-0.10	0.18	0.19
Fruits	-0.04	-0.03	-0.04	0.02	-0.08	-0.07	-0.01	-0.01	0.13	1.00	0.23	-0.03	0.03	-0.04	-0.09	-0.06	-0.04	-0.04	-0.09	0.07	0.09	0.07
Veggies	-0.07	-0.06	-0.04	-0.00	-0.05	-0.02	-0.04	-0.03	0.14	0.23	1.00	0.02	0.03	-0.03	-0.11	-0.05	-0.06	-0.08	-0.05	-0.02	0.14	0.15
HvyAlcoholConsump	-0.10	-0.03	-0.03	-0.03	-0.06	0.08	-0.02	-0.04	0.02	-0.03	0.02	1.00	-0.01	0.01	-0.06	0.01	-0.04	-0.05	0.02	-0.06	0.04	0.07
AnyHealthcare	0.03	0.04	0.03	0.11	-0.01	-0.01	0.01	0.02	0.02	0.03	0.03	-0.01	1.00	-0.22	-0.03	-0.05	-0.00	0.01	-0.01	0.14	0.10	0.13
NoDocbcCost	0.04	0.02	0.03	-0.06	0.06	0.03	0.03	0.03	-0.06	-0.04	-0.03	0.01	-0.22	1.00	0.16	0.19	0.15	0.12	-0.05	-0.13	-0.09	-0.19
GenHlth	0.40	0.31	0.23	0.06	0.26	0.14	0.19	0.27	-0.26	-0.09	-0.11	-0.06	-0.03	0.16	1.00	0.31	0.55	0.47	-0.02	0.15	-0.27	-0.37
MentHlth	0.08	0.06	0.08	-0.01	0.10	0.09	0.08	0.07	-0.12	-0.06	-0.05	0.01	-0.05	0.19	0.31	1.00	0.38	0.25	-0.09	-0.11	-0.10	-0.21
PhysHlth	0.21	0.17	0.14	0.04	0.16	0.11	0.16	0.19	-0.23	-0.04	-0.06	-0.04	-0.00	0.15	0.55	0.38	1.00	0.48	-0.05	0.08	-0.15	-0.27
DiffWalk	0.27	0.23	0.16	0.05	0.24	0.11	0.19	0.23	-0.27	-0.04	-0.08	-0.05	0.01	0.12	0.47	0.25	0.48	1.00	-0.08	0.19	-0.19	-0.34
Sex	0.04	0.04	0.01	-0.01	-0.00	0.11	0.00	0.10	0.05	-0.09	-0.05	0.02	-0.01	-0.05	-0.02	-0.09	-0.05	-0.08	1.00	-0.00	0.04	0.16
Age	0.27	0.33	0.24	0.10	-0.05	0.10	0.12	0.22	-0.10	0.07	-0.02	-0.06	0.14	-0.13	0.15	-0.11	0.08	0.19	-0.00	1.00	-0.10	-0.12
Education	-0.16	-0.13	-0.08	-0.01	-0.09	-0.13	-0.07	-0.09	0.18	0.09	0.14	0.04	0.10	-0.09	-0.27	-0.10	-0.15	-0.19	0.04	-0.10	1.00	0.45
Income	-0.21	-0.18	-0.10	0.01	-0.11	-0.09	-0.13	-0.14	0.19	0.07	0.15	0.07	0.13	-0.19	-0.37	-0.21	-0.27	-0.34	0.16	-0.12	0.45	1.00

Baseline Model:
Logistic Regression



Training Accuracy: 0.7401

Validation Accuracy: 0.7479

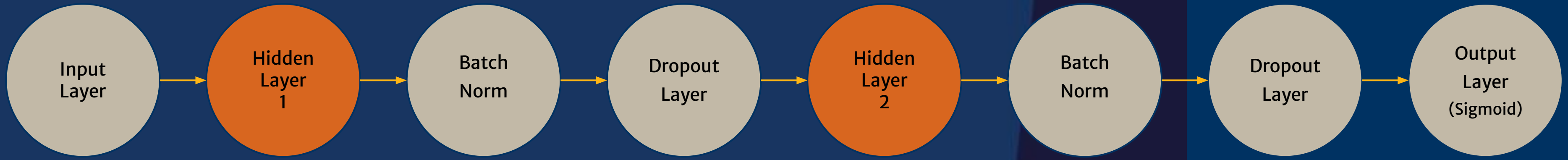
Testing Accuracy: 0.7349

Test Set - Diabetic & Pre-Diabetic Accuracy: 0.7617

Test Set - Non-Diabetic Accuracy: 0.7073

Improved Model

Neural Network



Berkeley
UNIVERSITY OF CALIFORNIA

Experiment

Learning Rate: 0.001; Batch Size: 64; Epoch: 200

Index	Hidden Size	Activation	Optimizer	Dropout Rate	Parameters	Training Accuracy	Validation Accuracy
1	[64, 32]	relu	SGD	[0.5, 0.3]	3139	0.7402	0.7508
2	[64, 32]	relu	Adam	[0.5, 0.3]	9029	0.7438	0.7503
3	[64, 32]	tanh	SGD	[0.5, 0.3]	3139	0.7405	0.7500
4	[64, 32]	tanh	Adam	[0.5, 0.3]	9029	0.7426	0.7503
5	[128, 64]	relu	SGD	[0.5, 0.3]	10371	0.7420	0.7501
6	[128, 64]	relu	Adam	[0.5, 0.3]	30341	0.7447	0.7514
7	[128, 64]	tanh	SGD	[0.5, 0.3]	10371	0.7407	0.7495
8	[128, 64]	tanh	Adam	[0.5, 0.3]	30341	0.7444	0.7506

Berkeley
UNIVERSITY OF CALIFORNIA

Experiment

Learning Rate: 0.001; Batch Size: 64; Epoch: 200

Index	Hidden Size	Activation	Optimizer	Dropout Rate	Parameters	Training Accuracy	Validation Accuracy
9	[128, 64]	relu	Adam	[0.3, 0.3]	30341	0.7443	0.7506
10	[128, 64]	relu	Adam	[0.4, 0.4]	30341	0.7450	0.7516
11	[128, 64]	relu	Adam	[0.5, 0.5]	30341	0.7447	0.7511
12	[128, 64]	relu	Adam	[0.4, 0.3]	30341	0.7456	0.7517
13	[128, 64]	relu	Adam	[0.3, 0.4]	30341	0.7446	0.7524
6	[128, 64]	relu	Adam	[0.5, 0.3]	30341	0.7442	0.7514
14	[128, 64]	relu	Adam	[0.3, 0.5]	30341	0.7449	0.7514
15	[128, 64]	relu	Adam	[0.4, 0.5]	30341	0.7449	0.7520
16	[128, 64]	relu	Adam	[0.5, 0.4]	30341	0.7447	0.7521

Berkeley
UNIVERSITY OF CALIFORNIA

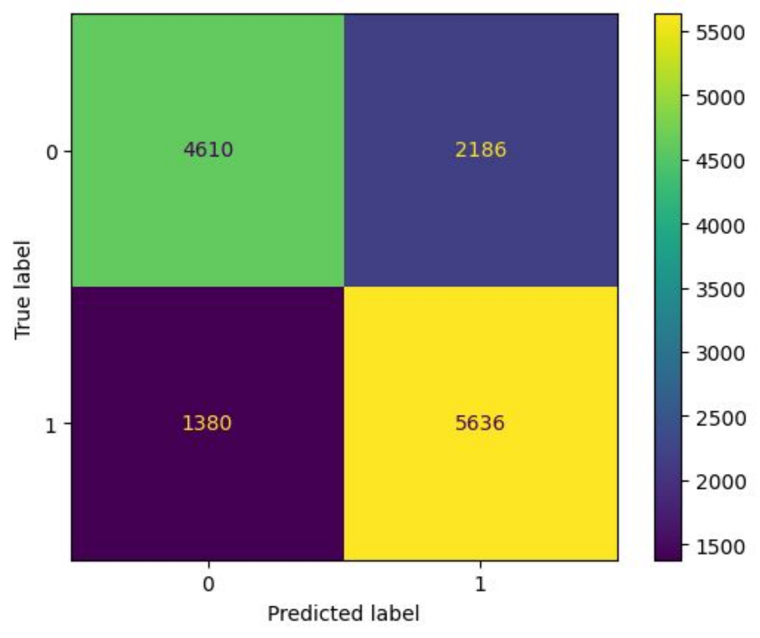
Experiment

Hidden Size: [128, 64]; Activation: relu; Optimizer: Adam; Dropout: [0.3, 0.4]

Index	Batch Size	Learning Rate	Epoch	Parameters	Training Accuracy	Validation Accuracy
17	128	0.001	200	30341	0.7442	0.7501
13	64	0.001	200	30341	0.7446	0.7524
18	32	0.001	200	30341	0.7447	0.7505
19	128	0.0001	200	30341	0.7467	0.7508
20	64	0.0001	200	30341	0.7450	0.7507
21	32	0.0001	200	30341	0.7458	0.7500
22	128	0.01	200	30341	0.7438	0.7530
23	64	0.01	200	30341	0.7436	0.7500
24	32	0.01	200	30341	0.7435	0.7501

Berkeley
UNIVERSITY OF CALIFORNIA

Conclusion - Result



Training Accuracy: 0.7438

Validation Accuracy: 0.7530

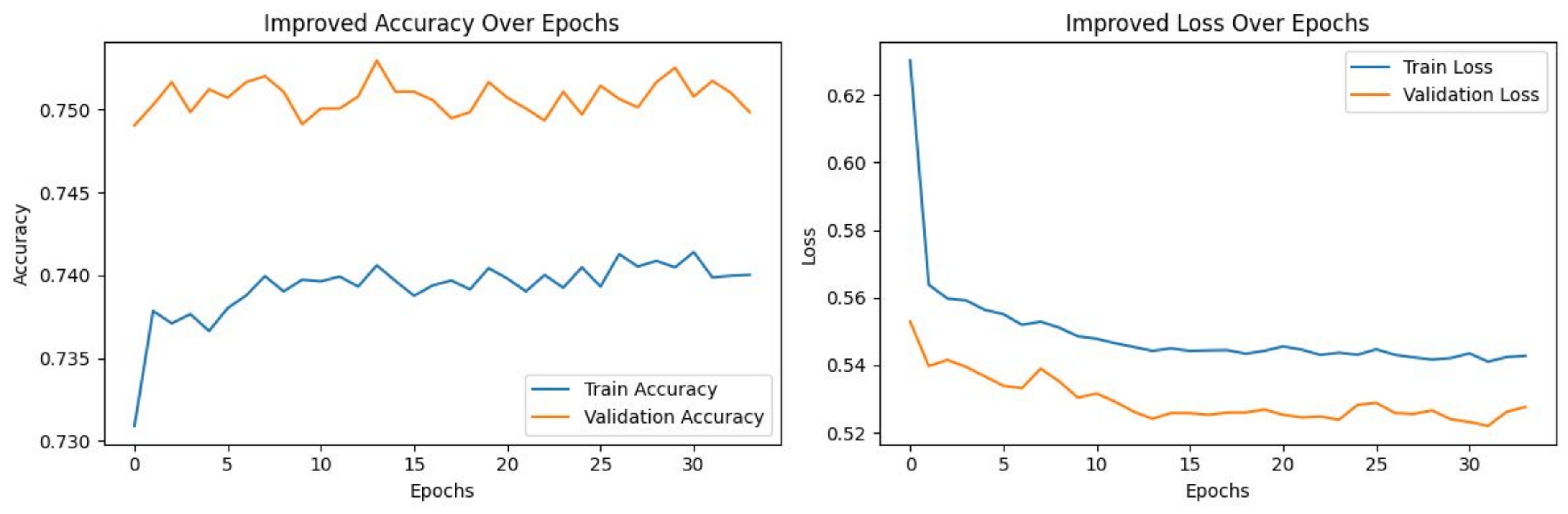
Testing Accuracy: 0.418

Test Set - Diabetic & Pre-Diabetic Accuracy: 0.8033

Test Set - Non-Diabetic Accuracy: 0.6783

Berkeley
UNIVERSITY OF CALIFORNIA

Conclusion - Result



Berkeley
UNIVERSITY OF CALIFORNIA

Conclusion - Avenue for Future

Considering more advanced model, for example: XGBoost, etc.

Exploring better hyperparameter tuning approaches, such as grid search

Looking for better datasets with potential for more feature engineering

Link to Github Repository

https://github.com/yuefengxue/mids-w207-final-YueFeng-Xue-Irina-Lee

Contribution

YueFeng Xue: Data preparation, data processing, EDA, modeling, evaluation & conclusion, design and draft PowerPoint, PowerPoint preparation, code clean-up and organize.

Irina Lee: Topic research and suggestion, data research and making a decision on what data we are using, modeling, evaluation & conclusion, code review, PowerPoint review and preparation.

Berkeley
UNIVERSITY OF CALIFORNIA

Thank You!

Berkeley
UNIVERSITY OF CALIFORNIA