

DSE-230

FINAL PROJECT PRESENTATION

DIABETES RISK PREDICTION FROM PERSONAL HEALTH INDICATORS

TEAM 3 Sergey Gurvich, Chunxia Tong, Camm Perera



AGENDA

- Problem definition
- Data description
- Data preparation
- Analysis approaches
- Analysis results
- Challenges and approaches to address challenges
- Insights gained
- Future work



PROBLEM DEFINITION

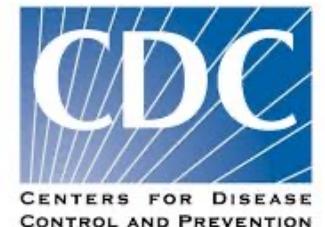
Background:

- Quick Fact: 37.3 million US adults have diabetes, and 1 in 5 of them don't know they have it. Diabetes is the seventh leading cause of death in the United States.
- Risk Factors: smoking, obesity, physical inactivity, family history, ethnicity.
- Health Complications: heart and kidney diseases, oral, vision, hearing problems, mental health.

Problem Definition:

- Problem: based on personal health indicators, predict if a person is at risk of developing the diabetes disease.
- Applications: healthcare and/or insurance industries.
- Success Criterion: build a binary classification model that can make a prediction with 85% accuracy rate.

DATA DESCRIPTION



Original Dataset:

- The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.
- Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories.
- BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

source: <https://www.cdc.gov/brfss/about/index.htm>

DATA DESCRIPTION



Original Dataset:

- Original Dataset Files, Scripts and Description: https://www.cdc.gov/brfss/annual_data/annual_2020.html
- 401,958 observations
 - Each sample represents individual's responses
- 279 features
 - Questionnaire responses
 - Calculated variables

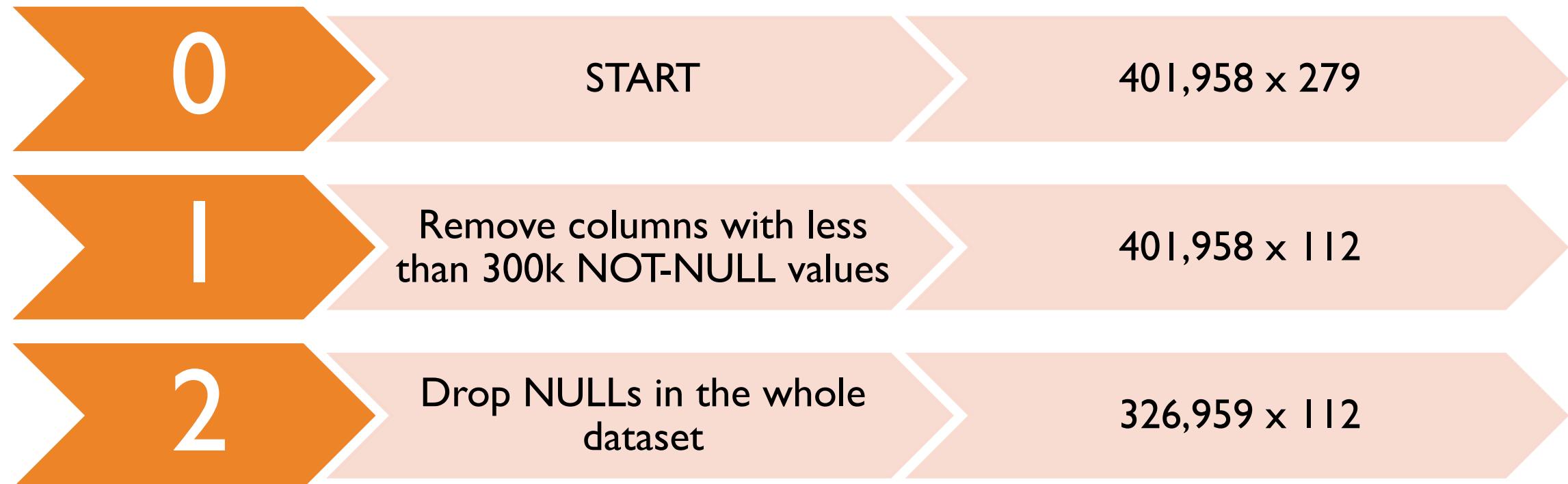
DATA DESCRIPTION



Data Examples:

Feature type	Feature Description	Possible Values
Question	GENHLTH: Would you say that in general your health is:	1=Excellent 2=Very good 3=Good 4=Fair 5=Poor 7=DK/NS 9=Refused
Question	EXERANY2: During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?	1=Yes 2>No 7=DK/NS 9=Refused
Calculated Variable	_PHYS14D: Calculated variable for 3 level not good physical health status.	1: 0 days, 2: 1-13 days, 3: 14-30 days, 9: Don't know/ Refused/ Missing

DATA PREPARATION



DATA PREPARATION

3

Performed EDA, found correlations with 'Diabetes' feature

4

Remove features with absolute correlation coefficient < 0.1

5

Remove highly correlated - linearly dependent features (based on CDC's provided documentation)

GENHLTH	EXERANY2	HAVARTH4	DIABETE4	EMPLOY1	DIFFWALK	ALCDAYS	_HCVU651	_MICHD	_DRDXAR2	_AGEG5YR	WTKG3	_BMIS	_BMISCAT
1.00	0.26	-0.29	-0.24	0.27	-0.30	0.18	0.14	-0.24	-0.29	0.19	0.16	0.22	0.19
0.26	1.00	-0.13	-0.12	0.14	-0.19	0.14	0.09	-0.10	-0.13	0.11	0.08	0.13	0.10
-0.29	-0.13	1.00	0.14	-0.31	0.24	-0.11	-0.29	0.17	1.00	-0.37	-0.06	-0.13	-0.12
-0.24	-0.12	0.14	1.00	-0.16	0.15	-0.13	-0.14	0.17	0.14	-0.18	-0.15	-0.19	-0.17
0.27	0.14	-0.31	-0.16	1.00	-0.23	0.19	0.56	-0.21	-0.31	0.51	-0.08	-0.01	-0.02
-0.30	-0.19	0.24	0.15	-0.23	1.00	-0.12	-0.13	0.15	0.24	-0.17	-0.08	-0.13	-0.10
0.18	0.14	-0.11	-0.13	0.19	-0.12	1.00	0.14	-0.09	-0.11	0.14	-0.00	0.07	0.05
0.14	0.09	-0.29	-0.14	0.56	-0.13	0.14	1.00	-0.21	-0.29	0.77	-0.12	-0.07	-0.04
-0.24	-0.10	0.17	0.17	-0.21	0.15	-0.09	-0.21	1.00	0.17	-0.23	-0.05	-0.05	-0.05
-0.29	-0.13	1.00	0.14	-0.31	0.24	-0.11	-0.29	0.17	1.00	-0.37	-0.06	-0.13	-0.12
0.19	0.11	-0.37	-0.18	0.51	-0.17	0.14	0.77	-0.23	-0.37	1.00	-0.07	-0.01	0.02
0.16	0.08	-0.06	-0.15	-0.08	-0.08	-0.00	-0.12	-0.05	-0.06	-0.07	1.00	0.86	0.75
0.22	0.13	-0.13	-0.19	-0.01	-0.13	0.07	-0.07	-0.05	-0.13	-0.01	0.86	1.00	0.84
0.19	0.10	-0.12	-0.17	-0.02	-0.10	0.05	-0.04	-0.05	-0.12	0.02	0.75	0.84	1.00

326,959 x 25

326,959 x 12

DATA PREPARATION

6

Went feature by feature, looked at the questions and responses:
Removed rows with 'Refused' response codes: 9/99/999
Removed outliers from BMI (BMI>60)
Replaced response codes: 2(No) by 0 for Logistic Regression / SVM Models
Replaced response codes: 8/88/888 by 0

312,703 x 12

7

Feature Transformation:
Scaling
Normalization
One-Hot Encoding to categorical features with more than 2 categories

THE END!



DATA PREPARATION

6

Went feature by feature, looked at the questions and responses:
Removed rows with 'Refused' response codes: 9/99/999
Removed outliers from BMI (BMI>60)
Replaced response codes: 2(No) by 0 for Logistic Regression / SVM Models
Replaced response codes: 8/88/888 by 0

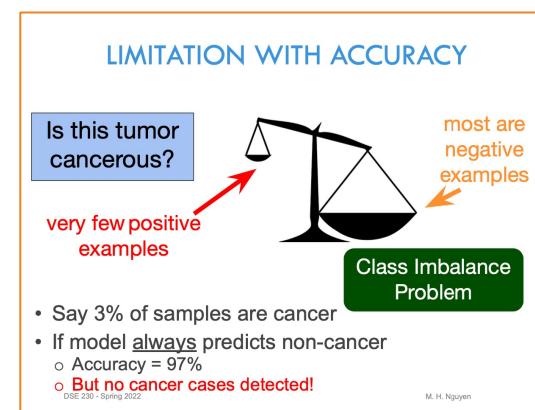
312,703 x 12

7

Feature Transformation:
Scaling
Normalization
One-Hot Encoding to categorical features with more than 2 categories



+-----+	-----+
label	count
+-----+	-----+
1	41786
0	270917
+-----+	-----+



Source: DSE230-S4-3-ModelEvaluation.pdf, Mai H. Nguyen

DATA PREPARATION

6

Went feature by feature, looked at the questions and responses:
Removed rows with 'Refused' response codes: 9/99/999
Removed outliers from BMI (BMI>60)
Replaced response codes: 2(No) by 0 for Logistic Regression / SVM Models
Replaced response codes: 8/88/888 by 0

$312,703 \times 12$

7

Feature Transformation:
Scaling
Normalization
One-Hot Encoding to categorical features with more than 2 categories

$312,703 \times 12$

8

Undersampling

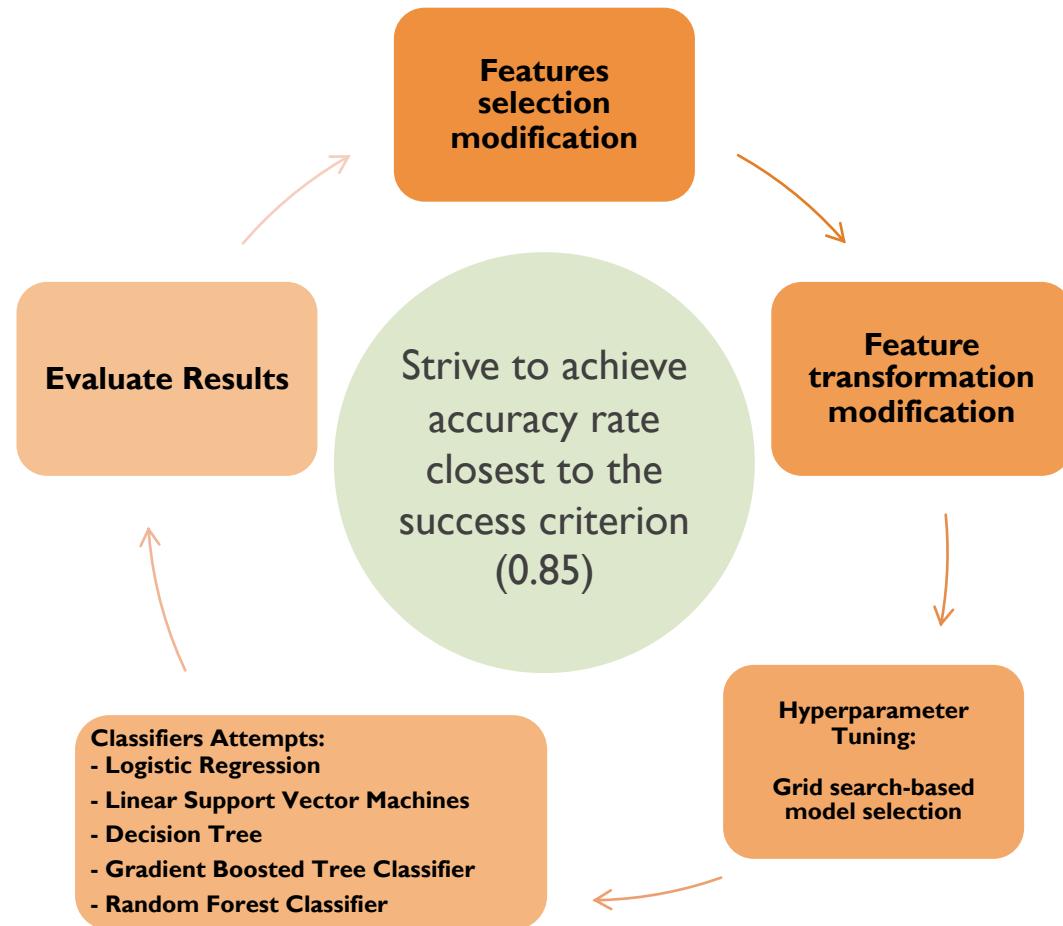
$83,555 \times 12$

ANALYSIS APPROACHES

Modeling (Iterative Process):

- Hyperparameter Tuning:

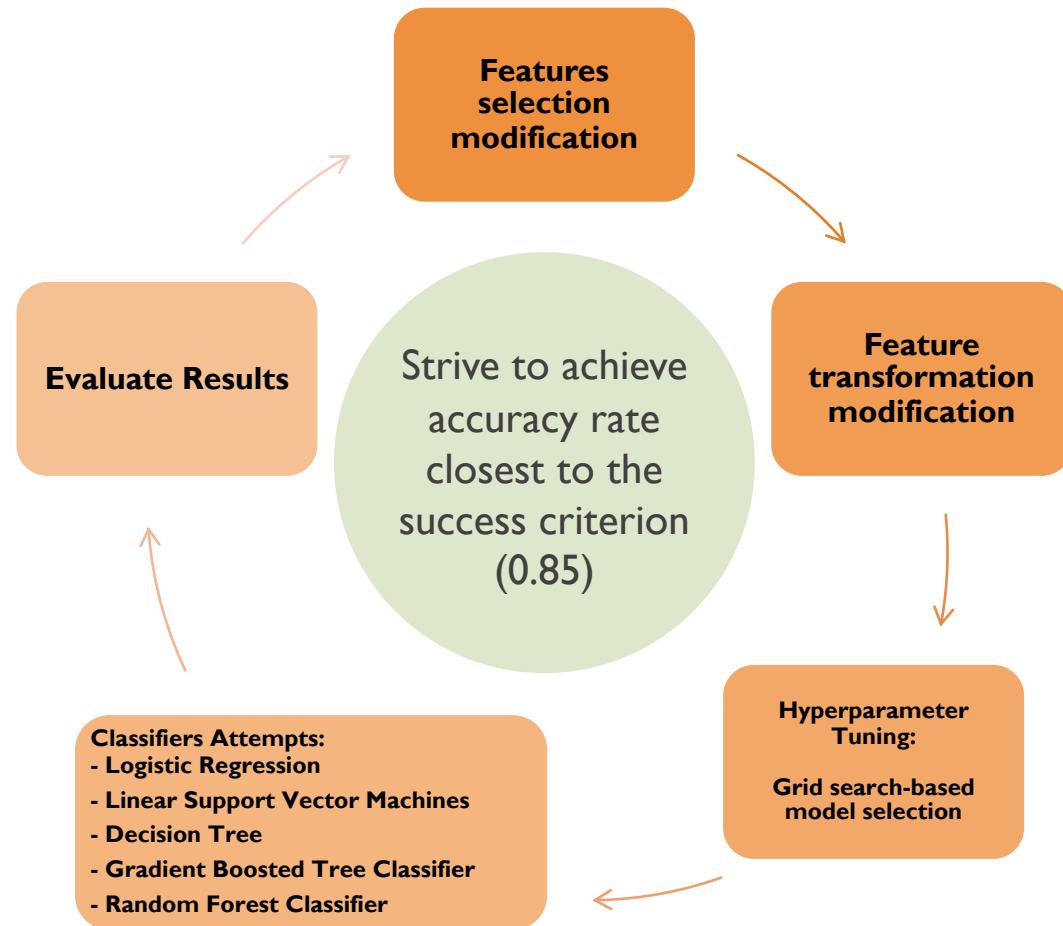
- Split the dataset into Train / Test (0.8 : 0.2)
- Applying K-Fold Cross Validation (K=10)



ANALYSIS APPROACHES

Modeling (Iterative Process):

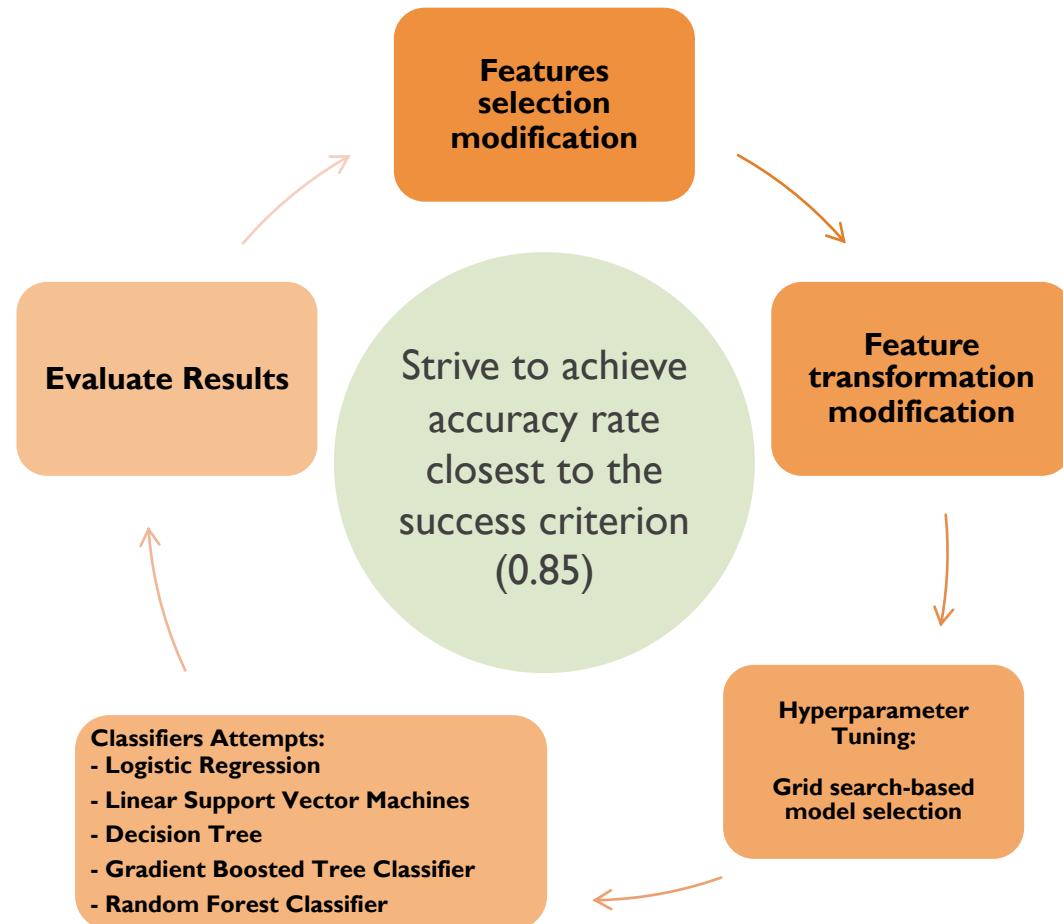
- Feature Engineering Modifications:
 - Cherry picking features
 - One-hot encoding
 - Standard Scaling
 - Vector Indexing



ANALYSIS APPROACHES

Modeling (Iterative Process):

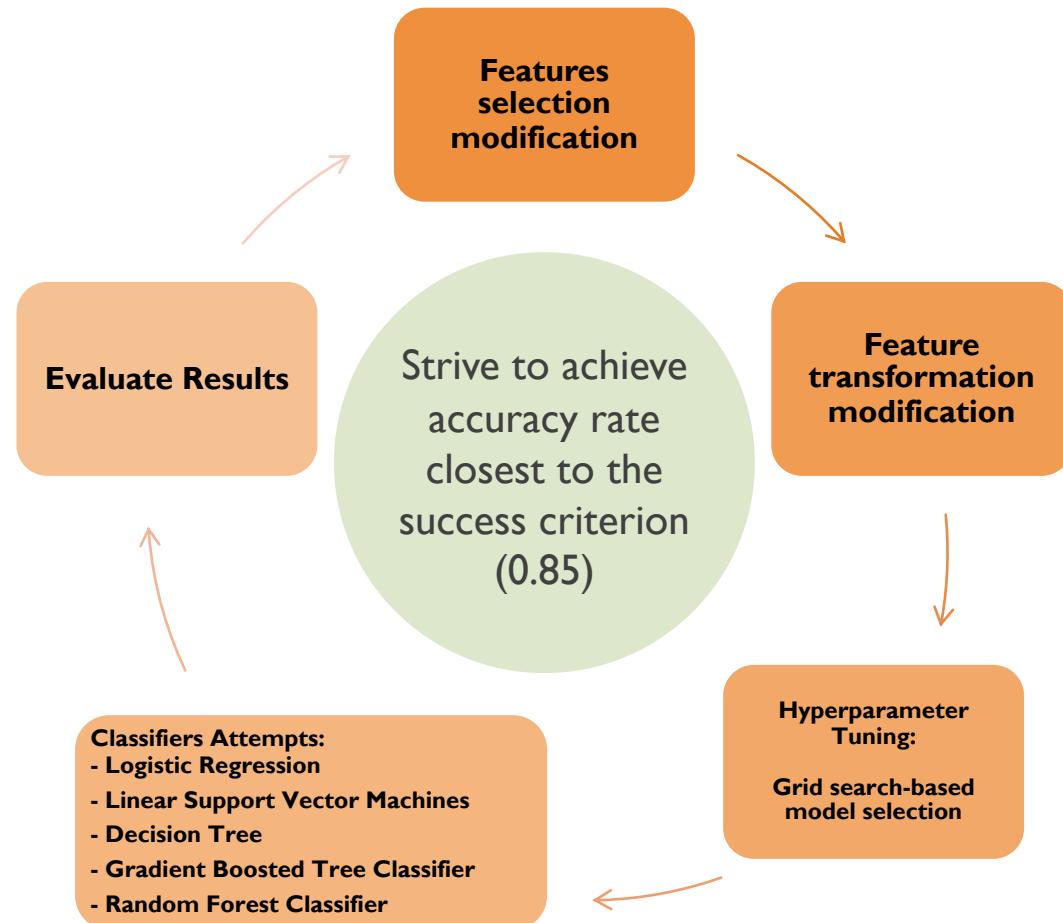
- Attempt different binary classification models:
 - Logistic Regression
 - Linear Support Vector Machines
 - Decision Tree
 - Gradient Boosted Tree Classifier
 - Random Forest Classifier



ANALYSIS APPROACHES

Modeling (Iterative Process):

- Model Evaluation:
 - Select the best model hyperparameters based on grid search cross validation results score (avgMetric)
 - Running training and test data through the selected best model and evaluation other metrics:
 - Accuracy / Test Error
 - Precision
 - Recall
 - F-Score
 - ROC Curve



ANALYSIS APPROACHES

Model Selection and Evaluation:

- Initially the model selection was done by comparing accuracy rates. But since we try to do the disease prediction, the problem requires specific evaluation approach (depending on the industry):
- Healthcare:
 - If somebody has risk of developing the disease, we want to catch it.
 - If somebody doesn't have that risk and we misclassified this case, it is better than failing in #1.
 - By combining #1 and #2, the best approach would be to try to maximize the RECALL metric, while we can afford some compromise on ACCURACY and PRECISION metrics



$$\text{ACCURACY} = \frac{\text{Correctly Predicted Individuals at Risk}}{\text{All Individuals}}$$



$$\text{PRECISION} = \frac{\text{Correctly Predicted Individuals at Risk}}{\text{All Predicted Individuals at Risk}}$$



$$\text{RECALL} = \frac{\text{Correctly Predicted Individuals at Risk}}{\text{Real Individuals at Risk}}$$

ANALYSIS APPROACHES

Model Selection and Evaluation:

- Initially the model selection was done by comparing accuracy rates. But since we try to do the disease prediction, the problem requires specific evaluation approach (depending on the industry):
- Insurance:
 - For insurance companies, the classification of negative cases is equally important as positive cases.
 - Therefore, the measure of exactness (PRECISION metric) would be more appropriate.



$$\text{ACCURACY} = \frac{\text{Correctly Predicted Individuals at Risk}}{\text{All Individuals}}$$



$$\text{PRECISION} = \frac{\text{Correctly Predicted Individuals at Risk}}{\text{All Predicted Individuals at Risk}}$$



$$\text{RECALL} = \frac{\text{Correctly Predicted Individuals at Risk}}{\text{Real Individuals at Risk}}$$

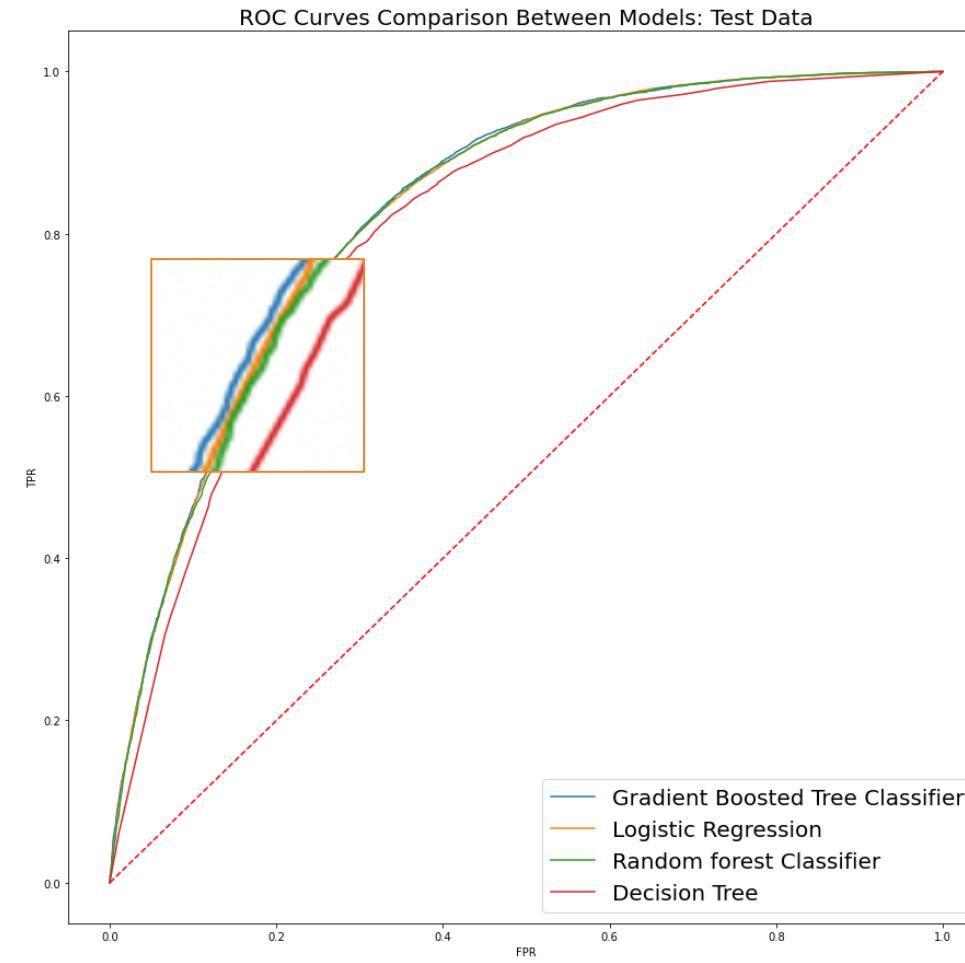
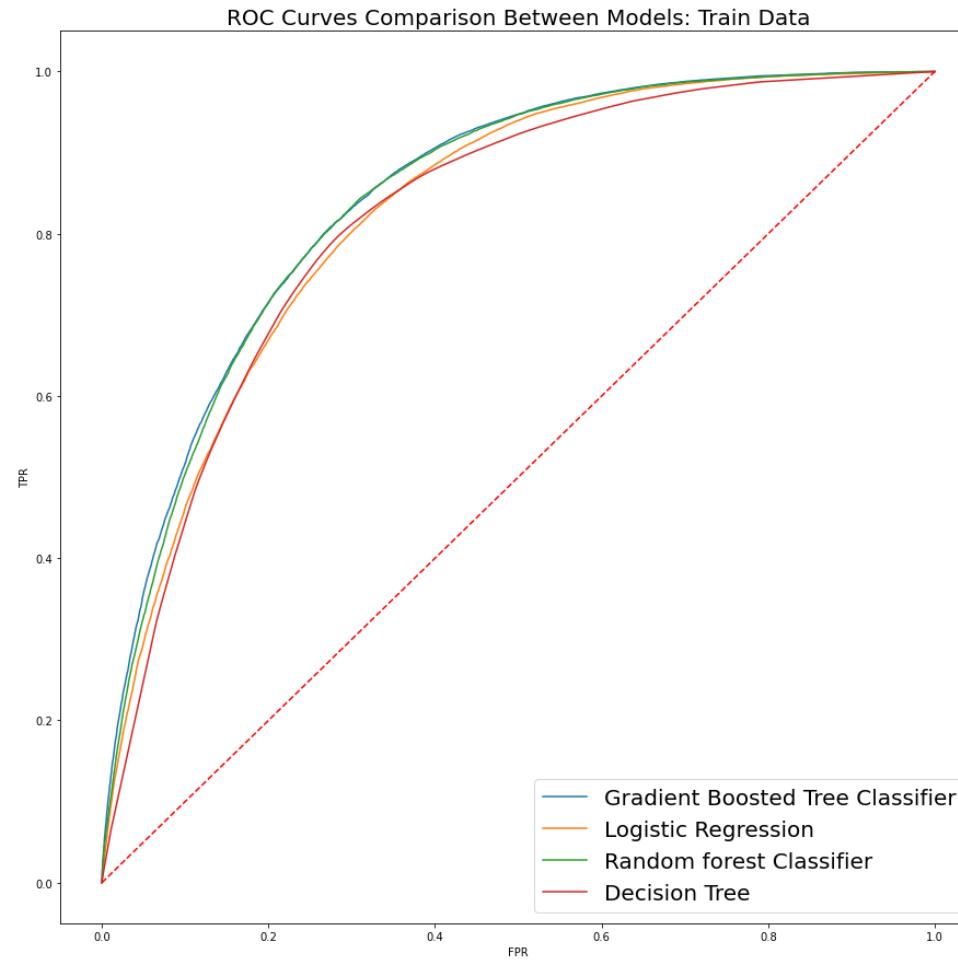
ANALYSIS RESULTS – MODELS' COMPARISON

won by a very narrow margin



					
	Logistic Regression	Linear SVM	Decision Tree	Gradient Boosted Tree	Random Forest
Accuracy (Train)	0.75	0.75	0.76	0.77	0.77
Train Error	0.25	0.25	0.24	0.23	0.23
Test Error	0.25	0.25	0.26	0.25	0.25
Precision (Test)	0.77	0.77	0.76	0.77	0.77
Recall (Test)	0.70	0.70	0.70	0.71	0.70
F-Score (Test)	0.74	0.73	0.73	0.74	0.74

ANALYSIS RESULTS – ROC CURVES



CHALLENGES AND APPROACHES TO ADDRESS CHALLENGES

General Challenges

- Selection out of 279 features (Feature Selection / Dimensionality Reduction):
 - ✓ use of common ML/DS techniques taught in the class
- Feature Transformation (which features required scaling / one-hot encoding):
 - ✓ consultation with experts
- Some functionalities of Pandas/Scikit Learn/ NumPy is not available in Spark:
 - ✓ developing custom code/functions
- Hyperparameter Tuning: time-constraints and level of experience:
 - ✓ online research
 - ✓ consultation with experts
 - ✓ use of common ML/DS techniques taught in the class

CHALLENGES AND APPROACHES TO ADDRESS CHALLENGES

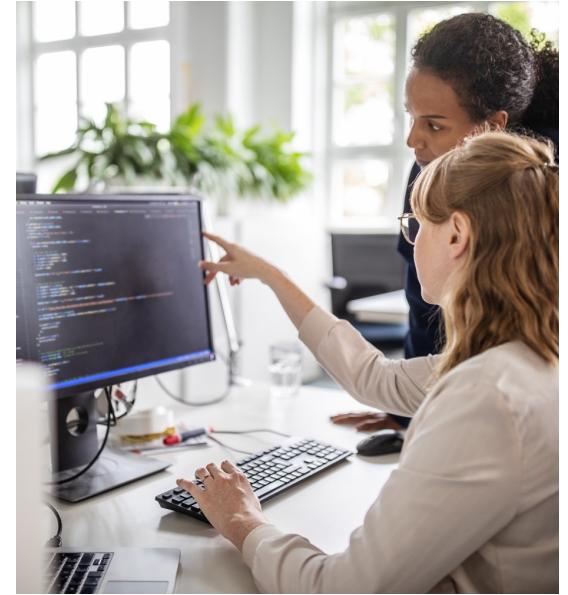
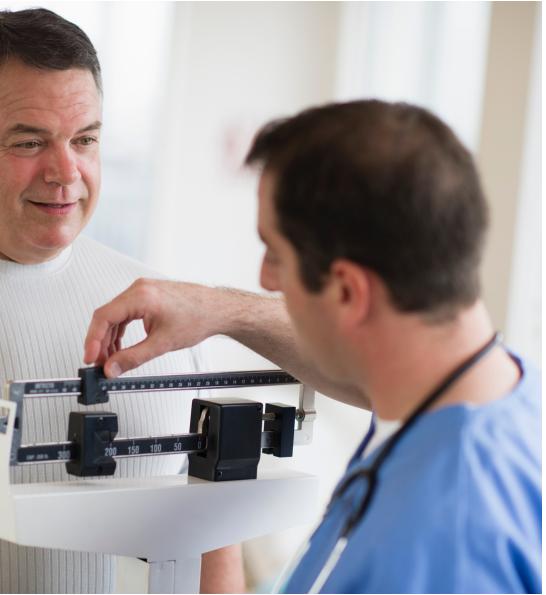
Specific Challenges

- The original dataset is in SAS format
 - ✓ Conversion to CSV in SAS Studio
- Imbalanced dataset
 - ✓ balancing by reducing the size of the majority class (under sampling)
- Dataset size reduced to 80k after balancing
 - ✓ Future work: combine with previous years data
- Feature Engineering: the dataset features has not uniform response codes (example: 9 vs 99 vs 999 for 'Refuse to answer')
 - ✓ Iterate through features to make them uniform
- Limited Computation Resources
 - ✓ Attempt to run in AWS EMR (failed)
 - ✓ Distributed tasks among team members
- Lack of Medical Domain Knowledge
 - ✓ used commonsense
 - ✓ online research
 - ✓ use of ML/DS techniques learned in class

INSIGHTS GAINED

- Health conditions such as Myocardial Infarction and Arthritis contribute to identifying diabetes risk in individuals.
- Feature selection based mainly on correlation coefficients might be not enough to achieve the accuracy goal of 0.85. (Future task can include supplementary datasets collected at state level).





FUTURE WORK

- Modelling/Evaluation Improvements:
 - Increase accuracy by more precise feature selection / engineering (using domain experts' knowledge)
 - Refine modeling process by combining historical and supplementary datasets (will expand sample size and relevant features count)
- Attempt to find probabilities of the risk
- Scale computational resources using the cloud platforms
- Expanding business use cases to other domains (marketing of medical devices)

QUESTIONS?
THANK YOU

