DSE-230

FINAL PROJECT PRESENTATION

DIABETES RISK PREDICTION FROM PERSONAL HEALTH INDICATORS

TEAM 3 Chunxia Tong, Camm Perera, Sergey Gurvich

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

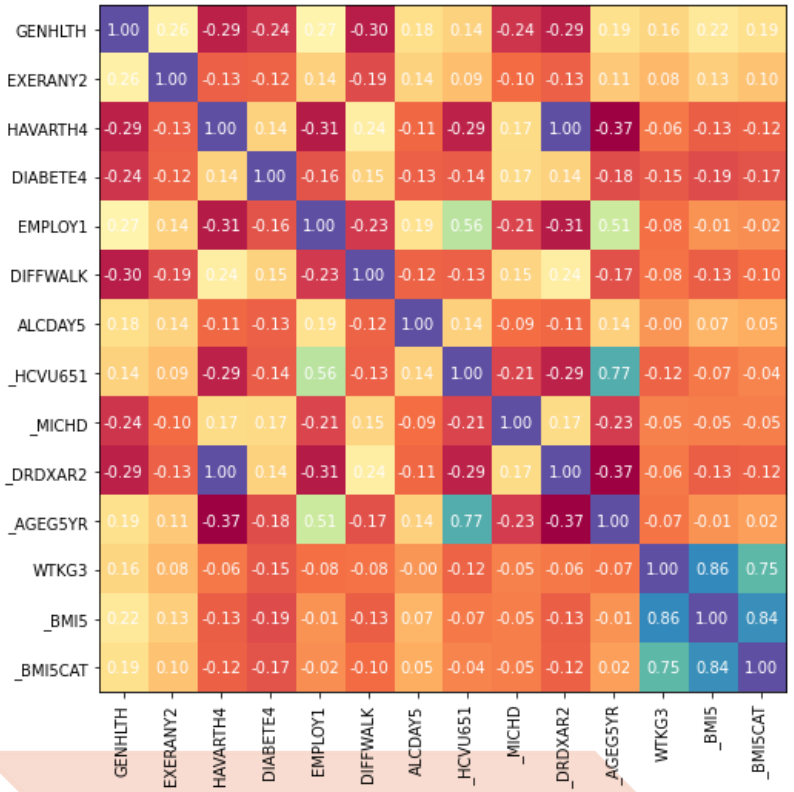
Data Examples:

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

0	START	401,958 x 279
1	Remove columns with less than 300k NOT-NULL values	401,958 x 112
2	Drop NULLs in the whole dataset	326,959 x 112

DATA PREPARATION



3 Performed EDA, found correlations with 'Diabetes' feature

4 Remove features with absolute correlation coefficient < 0.1

326,959 x 25

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

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

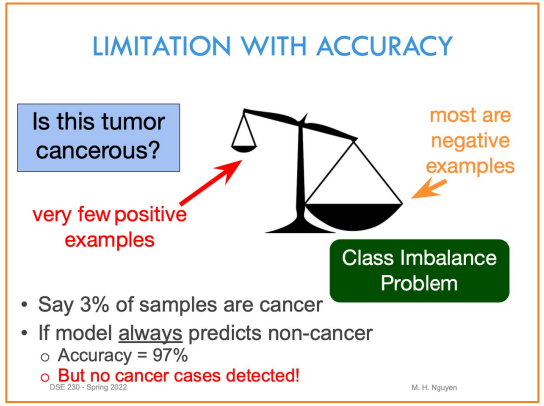
Went feature by feature, looked at the questions and responses:
Removed rows with 'Refused' response codes: 9/99/999
Removed outliers from BMI (BMI>60)
Replaced response codes: 2(No) by 0 for Logistic Regression / SVM Models
Replaced response codes: 8/88/888 by 0

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

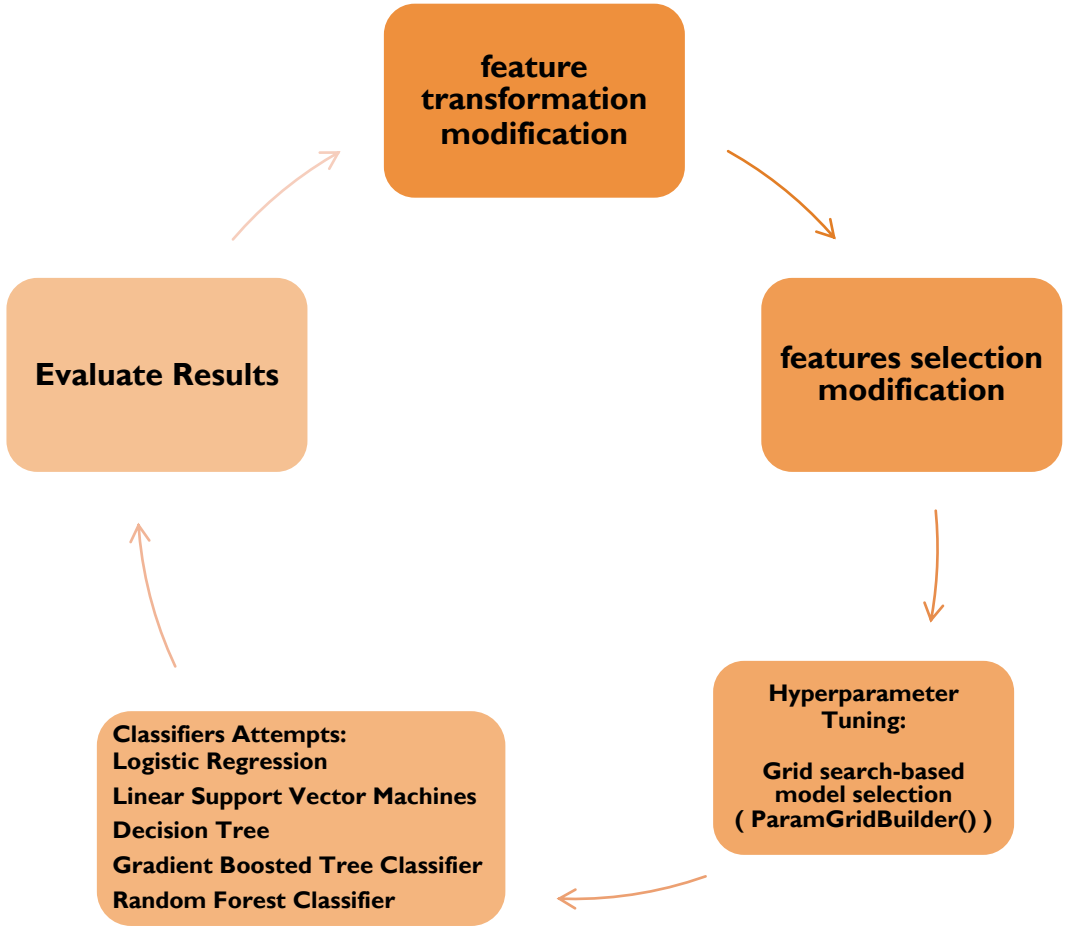
DATA PREPARATION

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 x 12
7	Feature Transformation: Scaling Normalization One-Hot Encoding to categorical features with more than 2 categories	312,703 x 12
8	Undersampling	83,555 x 12

ANALYSIS APPROACHES

Modeling:

- Split the dataset into Train / Test (0.8 : 0.2)
- Applying K-Fold Cross Validation (K=10)
- Striving to achieve accuracy rate closest to the success criterion (0.85)



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:
  - 1. If somebody has risk of developing the disease, we want to catch it.
  - 2. If somebody doesn't have that risk and we misclassified this case, it is better than failing in #1.
  - 3. 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.
  - 4. If comparing the ROC curves, the accent should be made on maximizing the True Positive Rate (TPR).

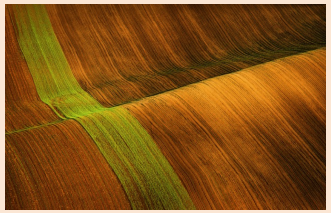
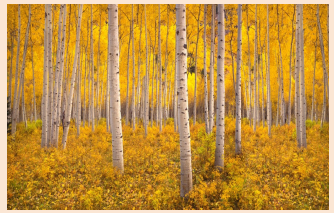
$$ACCURACY = \frac{Correctly\ Predicted\ Individuals\ at\ Risk}{All\ Individuals}$$

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

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

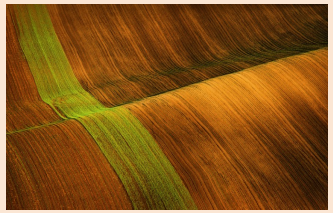
ANALYSIS RESULTS – MODELS' COMPARISON

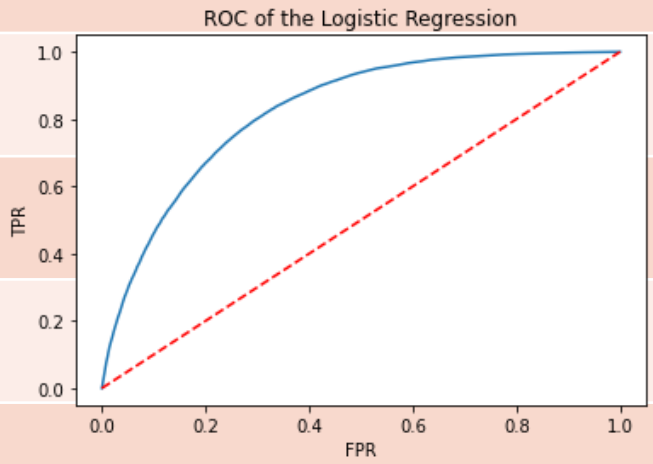
Best balance and high 'recall'

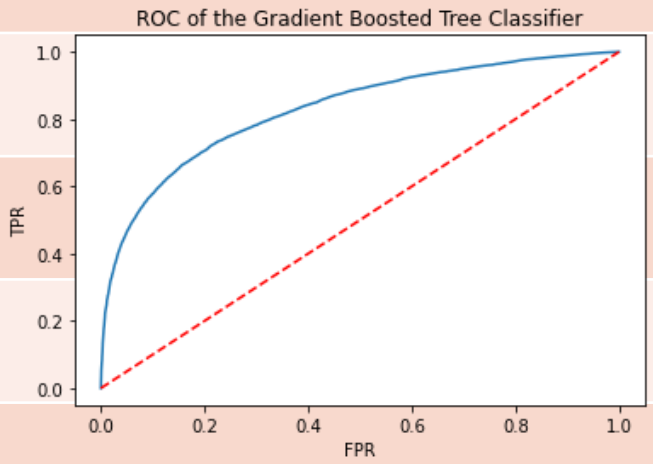
					
	Logistic Regression	Linear SVM	Decision Tree	Gradient Boosted Tree	Random Forest
Accuracy	0.75	0.75	0.72	0.75	0.74
Test Error	0.25	0.25	0.28	0.25	0.26
Precision	0.76	0.77	0.68	0.74	0.72
Recall	0.72	0.70	0.84	0.80	0.78

ANALYSIS RESULTS – MODELS' COMPARISON

Best balance and high 'recall'

		
	Logistic Regression	Gradient Boosted Tree
Accuracy	0.75	0.75
Test Error	0.25	0.25
Precision	0.76	0.74
Recall	0.72	0.80





CHALLENGES AND APPROACHES TO ADDRESS CHALLENGES

General Challenges

- Selection out of 279 features (Feature Selection / Dimensionality Reduction)
- 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
- Model Hyperparameter Fine-Tuning: time-constrains / lack of expertise
  - online research
  - consultation with experts
  - use of common ML/DS techniques taught in the class

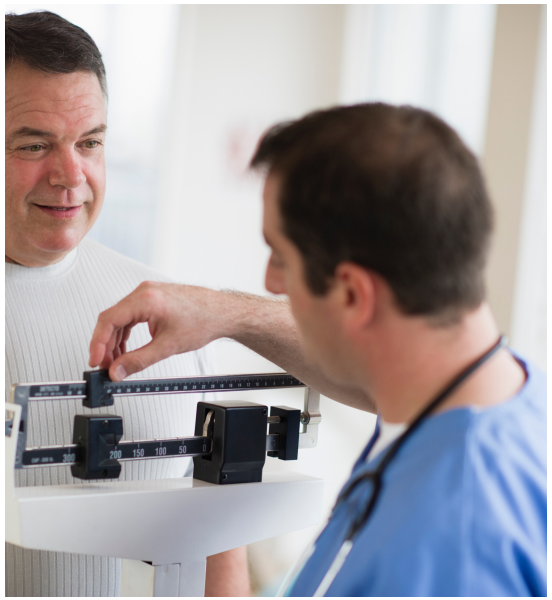
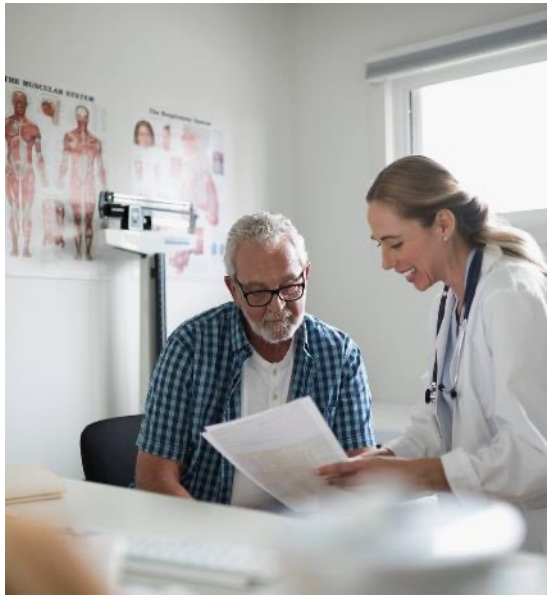
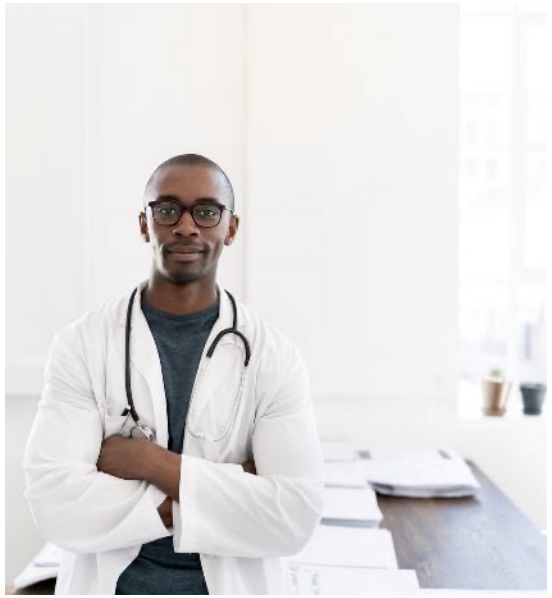
CHALLENGES AND APPROACHES TO ADDRESS CHALLENGES

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 (undersampling)
- 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')
  - had to go feature by feature to make it uniform
- Lack of Medical Domain Knowledge
  - use of commonsense, online research
  - use of common ML/DS techniques taught in the class

INSIGHTS GAINED

- For this type of problems decision-tree based models perform better
- Feature Selection insights (TBD)


FUTURE WORK

- Increase accuracy by more precise feature selection / engineering
- Discussions with domain experts
- Evaluation of models with combined historical datasets
- Models hyperparameters fine-tuning continuation
- Attempt to find probabilities for the results

QUESTIONS?

THANK YOU