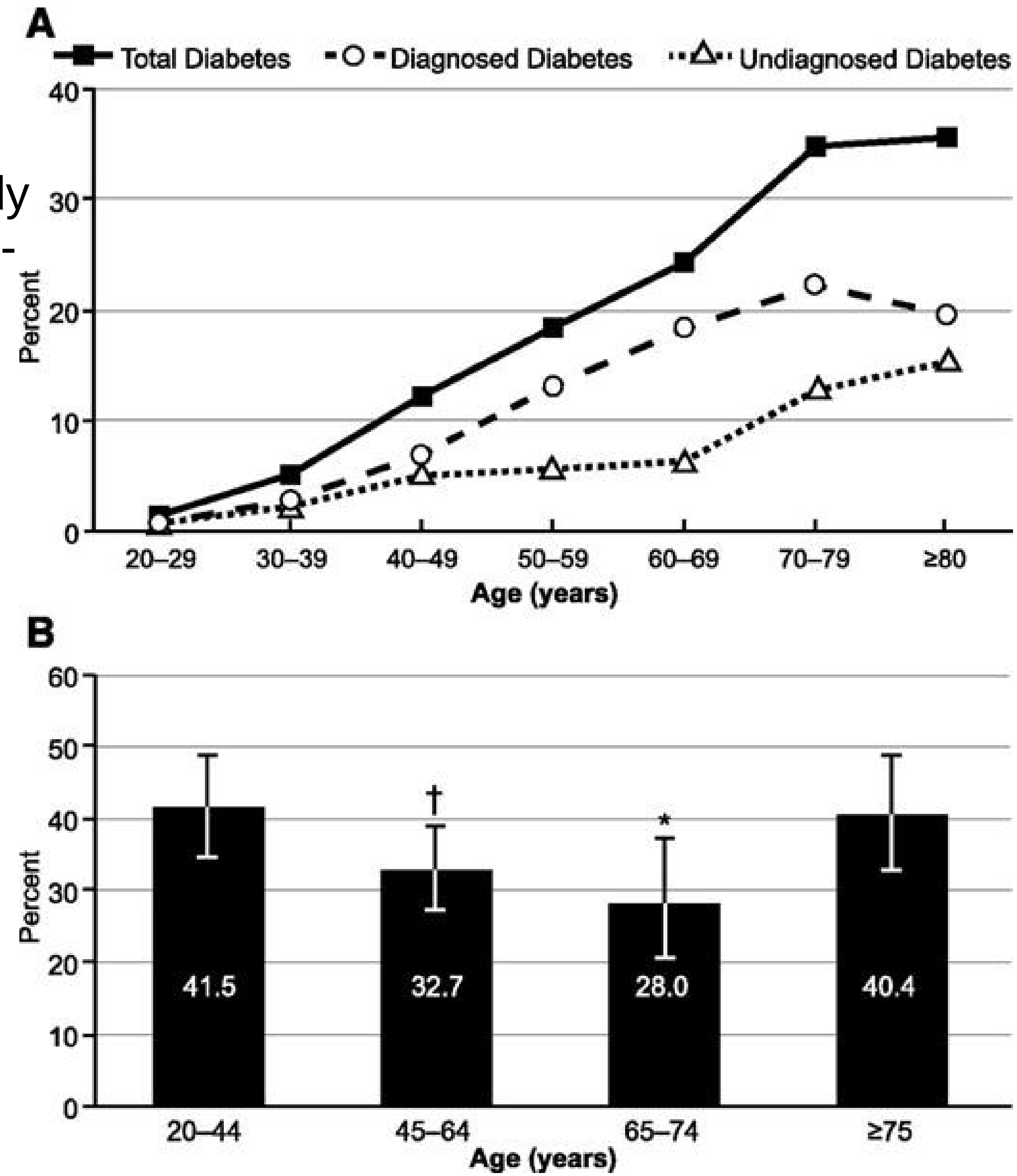
Predicting Diabetes from Telephone Interviews

Joe Hardin
Introduction to Machine Learning: Supervised Learning
University of Colorado – Boulder

What is Diabetes

- Diabetes is due to either the pancreas not producing enough insulin, or the cells of the body not responding properly to the insulin produced - Wikipedia
- In 2019, diabetes resulted in approximately 4.2 million deaths. It is the 7th leading cause of death globally. - Wikipedia
- Learning about the disease and actively participating in the treatment is important, since complications are far less common and less severe in people who have well-managed blood sugar levels. - Wikipedia



https://en.wikipedia.org/wiki/Diabetes
Graph source: https://diabetesjournals.org/care/article/42/6/994/35979/Diabetes-Diagnosis-and-Control-Missed

BRFSS

Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

Behavioral Risk Factor Surveillance System

CDC > BRFSS

- BRFSS
- About BRFSS
  - BRFSS Today
  - BRFSS History
  - BRFSS FAQs
- Archived
- Prevalence Data and Data Analysis Tools
- Survey Data and Documentation

About BRFSS

Print

BRFSS™ 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.

By collecting behavioral health risk data at the state and local level, BRFSS has become a powerful tool for targeting and building health promotion activities. As a result, BRFSS users have increasingly demanded more data and asked for more questions on the survey. Currently, there is a wide sponsorship of the BRFSS survey, including most divisions in the CDC National Center for Chronic Disease Prevention and Health Promotion; other CDC centers; and federal agencies, such as the Health Resources and Services Administration, Administration on Aging, Department of Veterans Affairs, and Substance Abuse and Mental Health Services Administration.

Screenshot from: https://www.cdc.gov/brfss/about/index.htm

Purpose

Can we identify individuals at high risk for having diabetes from the answers given in the BRFSS?

Actual Values

Predicted Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
https://www.tutorialexample.com/an-introduction-to-accuracy-precision-recall-f1-score-in-machine-learning-machine-learning-tutorial/

BRFSS Continued

- 2015 Phone Script can be found here : https://www.cdc.gov/brfss/questionnaires/pdf-ques/2015-brfss-questionnaire-12-29-14.pdf

- Potential sources of bias:

  - Only interviewed people with landlines in their residence (potential to bias by age, housing security, other factors)

  - Only conducted in English (may under count certain populations in the USA)

  - Methodology selects for people who answer unknown numbers and have the time to answer 20-58 questions (less clear if this sub population is more prevalent in some groups compared to others)

  - People's perception of the truth, not necessarily an objective TRUTH

Data Exploration

CDC's BRFSS – Original Source

https://www.cdc.gov/brfss/about/index.htm

CENTERS FOR DISEASE CONTROL AND PREVENTION · UPDATED 5 YEARS AGO

154

Behavioral Risk Factor Surveillance System

Public health surveys of 400k people from 2011-2015

Full Dataset uploaded by the CDC to Kaggle

ALEX TEBOUL · UPDATED A YEAR AGO

Diabetes Health Indicators Dataset

253,680 survey responses from cleaned BRFSS 2015 + balanced dataset

Alex Teboul put uploaded a clean dataset with all the responders in 2015 who answered all core questions to Kaggle

Data Exploration Continued

Core Sections 8

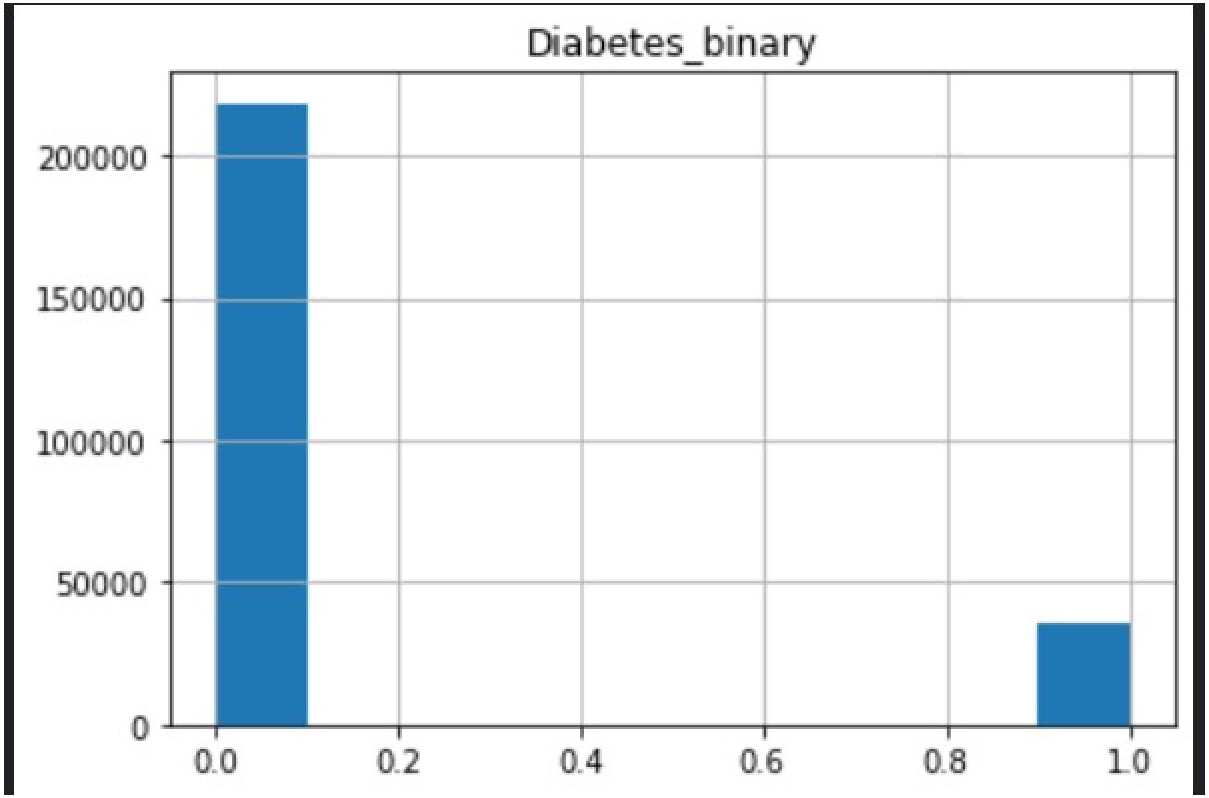
Section 1: Health Status 8
Section 2: Healthy Days — Health-Related Quality of Life 8
Section 3: Health Care Access 9
Section 4: Hypertension Awareness ....
Section 5: Cholesterol Awareness ....
Section 6: Chronic Health Conditions ....
Section 7: Demographics ....
Section 8: Tobacco Use ....
Section 9: Alcohol Consumption ....
Section 10: Fruits and Vegetables ....
Section 11: Exercise (Physical Activity) ....
Section 12: Arthritis Burden ....
Section 13: Seatbelt Use ....
Section 14: Immunization ....
Section 15: HIV/AIDS ....

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
```

y:Binary(#0)

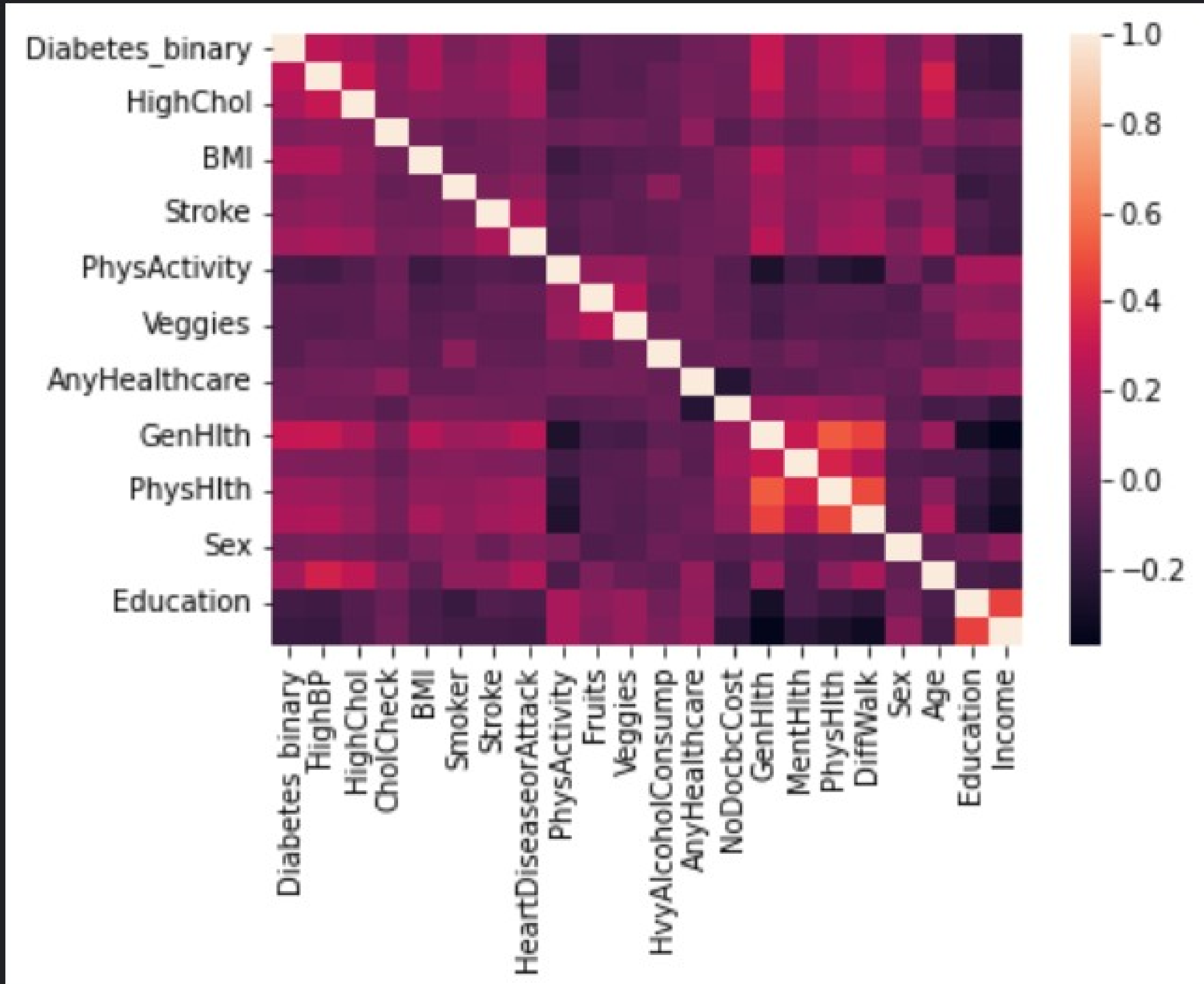
X : 21 features (#1-21)

Data Exploration Continued



- Unbalanced Data
- 0 – Non Diabetic
- 1 – Prediabetic/Diabetic
- BFRSS was trinary
- ~16% Positive

Data Exploration Continued

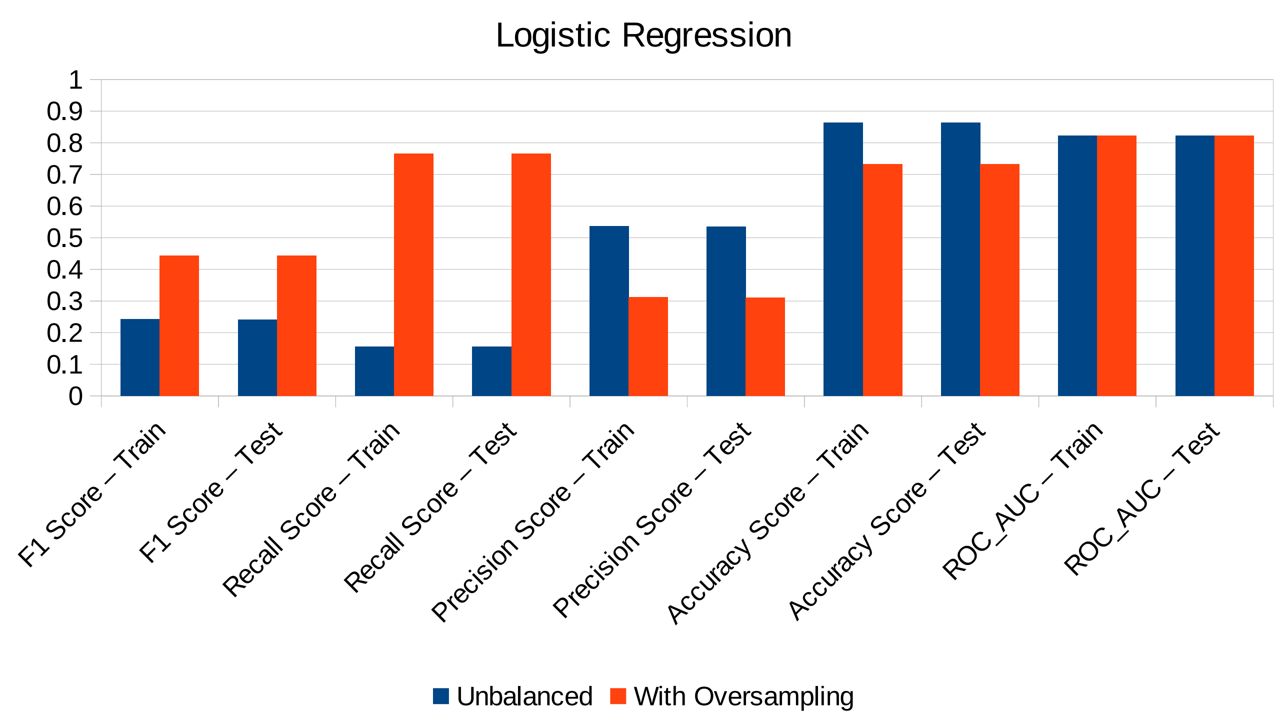


Supervised Machine Learning Methodology

```
%%time
#Cell runtime --  45 seconds

steps = [('over', RandomOverSampler()), ('model', DecisionTreeClassifier())]
pipeline = Pipeline(steps=steps)
# evaluate pipeline
cv = RepeatedStratifiedKFold(n_splits=5, n_repeats=2)
scores = pd.DataFrame.from_dict( cross_validate(pipeline, data[x], data[y], cv=cv,
                          scoring=('f1',
                                   'recall' ,
                                   'precision',
                                   'precision_micro',
                                   'accuracy',
                                   'roc_auc'),
                         return_train_score=True))

scores = scores.mean(axis=0)
pp.pprint(scores)
```

Logistic Regression

	Unbalanced	With Oversampling
F1 Score – Train	0.242	0.433
F1 Score – Test	0.241	0.433
Recall Score – Train	0.156	0.766
Recall Score – Test	0.156	0.766
Precision Score – Train	0.536	0.312
Precision Score – Test	0.535	0.311
Accuracy Score – Train	0.864	0.732
Accuracy Score – Test	0.863	0.732
ROC_AUC – Train	0.822	0.823
ROC_AUC – Test	0.822	0.823



- Oversampling Increased Recall decreased Precision
- Both Unbalanced and Oversampled data were scaled before model was trained.

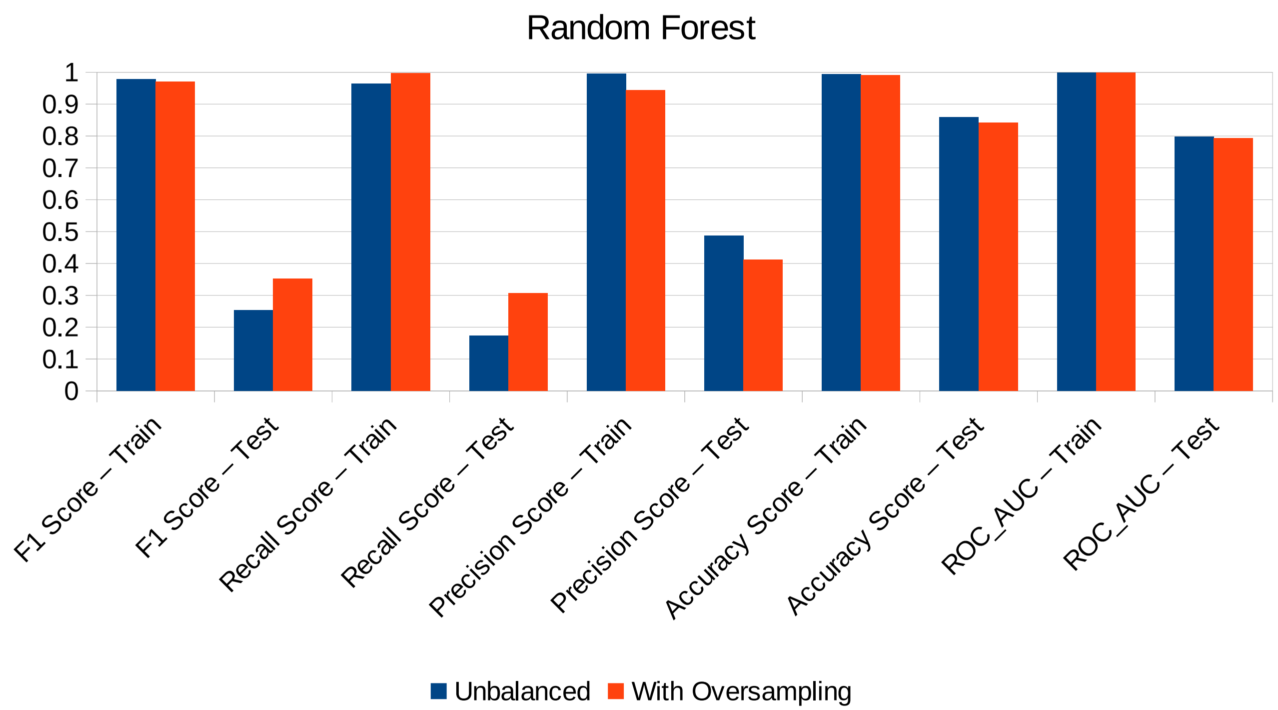
Single Decision Tree

	Unbalanced	With Oversampling
F1 Score – Train	0.979	0.971
F1 Score – Test	0.307	0.291
Recall Score – Train	0.960	0.997
Recall Score – Test	0.323	0.298
Precision Score – Train	0.999	0.946
Precision Score – Test	0.293	0.285
Accuracy Score – Train	0.994	0.992
Accuracy Score – Test	0.797	0.798
ROC_AUC – Train	0.999	0.999
ROC_AUC – Test	0.598	0.589

- Clearly Overfit – Further Tuning Required

Random Forest

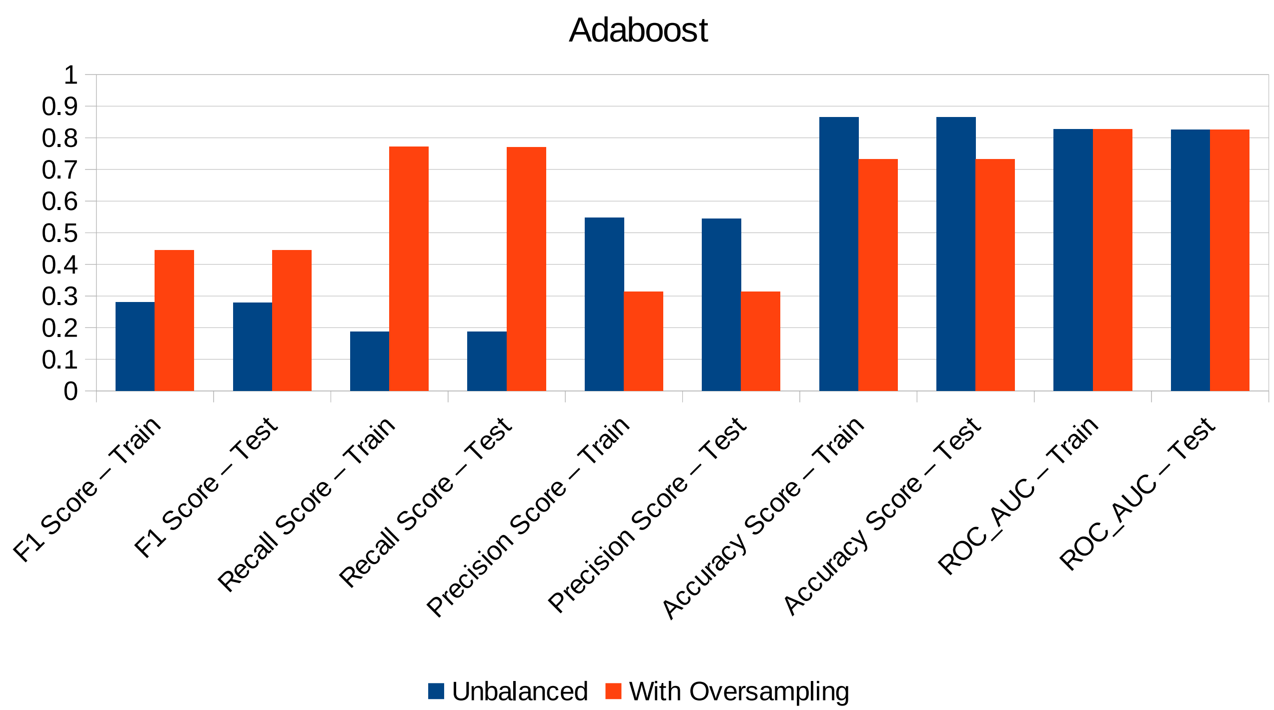
	Unbalanced	With Oversampling
F1 Score – Train	0.979	0.97
F1 Score – Test	0.254	0.352
Recall Score – Train	0.964	0.997
Recall Score – Test	0.173	0.307
Precision Score – Train	0.995	0.944
Precision Score – Test	0.487	0.412
Accuracy Score – Train	0.994	0.991
Accuracy Score – Test	0.859	0.842
ROC_AUC – Train	0.999	0.999
ROC_AUC – Test	0.798	0.793



- Clearly Overfit – Further Tuning Required

AdaBoost

	Unbalanced	With Oversampling
F1 Score – Train	0.28	0.445
F1 Score – Test	0.279	0.445
Recall Score – Train	0.188	0.772
Recall Score – Test	0.187	0.771
Precision Score – Train	0.547	0.313
Precision Score – Test	0.544	0.313
Accuracy Score – Train	0.865	0.732
Accuracy Score – Test	0.865	0.732
ROC_AUC – Train	0.827	0.827
ROC_AUC – Test	0.826	0.826

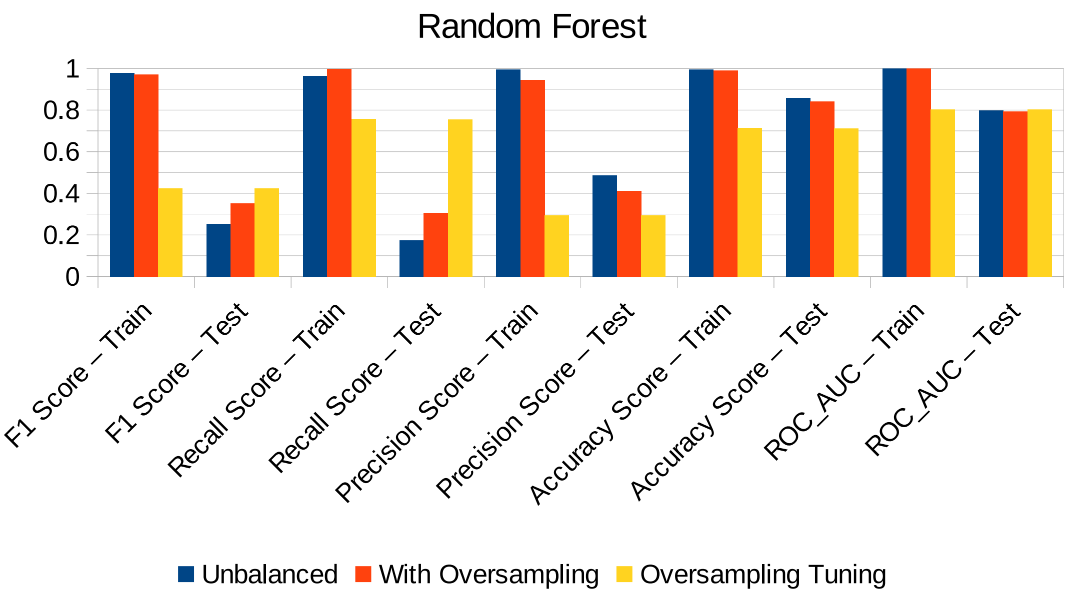


Tuning Random Forest Model

```
depths = range(2,5)
alphas =  np.logspace(-2,0,10)
cv = RepeatedStratifiedKFold(n_splits=5, n_repeats=1)
manualGrid = pd.DataFrame(columns = ['Depth', 'Alpha' , 'Scores'])

for depth in depths:
    for alpha in alphas:
        steps = [('over', RandomOverSampler()), ('model',RandomForestClassifier(max_depth = depth, ccp_alpha =
        pipeline = Pipeline(steps=steps)


        scores = pd.DataFrame.from_dict( cross_validate(pipeline, data[x], data[y], cv=cv,
                          scoring=('f1',
                                  'recall' ,
                                  'precision',
                                  'precision_micro',
                                  'accuracy',
                                  'roc_auc'),
                         return_train_score=True))

        scores = scores.mean(axis=0)
        print( " Depth : " , depth , " Alpha : " , alpha)
        manualGrid[len(manualGrid)] = [depth, alpha, scores]
        pp.pprint(scores)
```

Random Forest – Revisited

	Unbalanced	With Oversampling	Oversampling Tuning
F1 Score – Train	0.979	0.97	0.423
F1 Score – Test	0.254	0.352	0.423
Recall Score – Train	0.964	0.997	0.757
Recall Score – Test	0.173	0.307	0.756
Precision Score – Train	0.995	0.944	0.294
Precision Score – Test	0.487	0.412	0.293
Accuracy Score – Train	0.994	0.991	0.713
Accuracy Score – Test	0.859	0.842	0.712
ROC_AUC – Train	0.999	0.999	0.803
ROC_AUC – Test	0.798	0.793	0.802



Conclusion

Can we identify individuals at high risk for having diabetes from the answers given in the BRFSS?

- The best model tried (Adaboost with Oversampling) give the below results (cv=3)
- Depends on Stakeholder feedback

Actual Values

Predicted Values	Positive (1)	Negative (0)
Positive (1)	TP 27,240	FP 59820
Negative (0)	FN 8,106	TN 158,514

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$