Report on Diabetes

After importing the requisite packages, let us load our data and merge it as needed.

Additionally, let us describe the data set in basic terms

```
 (5000, 38)
ordinal data has 5000 rows and 11 columns 10 excluding PERSONID

 categorical data has 5000 rows and 23 columns 22 excluding PERSONID

 numeric_data has 5000 rows and 6 columns 5 excluding PERSONID
```

It appears we have loaded the data correctly, with the number of rows being equal across all our datasets as well as our final dataset

However, upon inspection of the columns we have several duplicates of the 'DIABETE3' column let us begin the process of cleaning our data by dropping these repeat columns

Before doing any more data cleaning let us get an idea of some of our data variables using some visualizations of our data. Firstly, let's get a visualization of the Diabetes category as it will be our primary response variable. Secondly, let's get a visualization of some of the variables that could be primary indicators in our model in our model. Let's focus on variables that are health income adjacent for our initial analysis as well as demographic data to ensure our sample is reasonably representative and not demographically biased

Health Variables:

- General Health: ordinal data so let us get a frequency chart
- Asthma
- Mental Health: numeric data let us rescale our data so that no days with poor mental health are encoded as zero

and values 'Not Sure', 'Blank', and 'Refused' are encoded as null

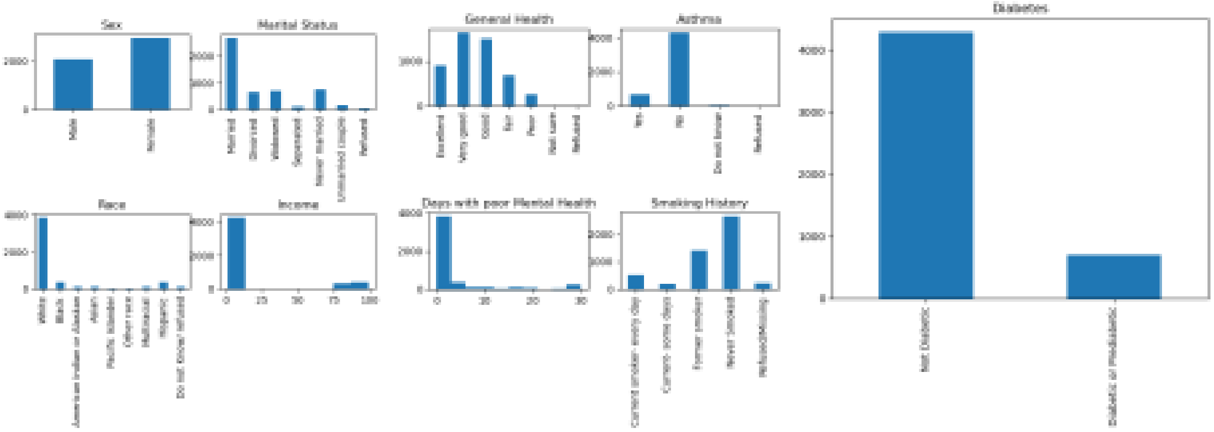
- Smoker

Demographic Variables

- Sex
- Marital Status
- Race
- Income: Like mental health numeric data let us rescale our data so that 'Blank', and 'Refused' are encoded as null

Finally Diabetes, this will be the response variable of our model.

```
Out[4]: <function matplotlib.pyplot.show(close=None, block=None)>
```



Data Cleaning

Categorical Data

Something we noticed is that blank and unsure responses can be recorded and regrouped together as they represent a very similar response. The advantage to this is reducing the number of categories reduces the chance of conflating variables or overfitting especially with these categories as they represent a small proportion

of responses We will only be doing this modification for variables that we select to use in the model as needed because the process is variable for each category.

Numeric Data

An issue that arises with numeric data is missing or refused values. Most prediction algorithms are not suited for missing values. To deal with this we can bin our data and group missing or refused data into a category together. Transforming the numeric data into categorical data. Income is already binned so we simply add a category for missing/refused encoded as 0.

Days with poor mental health has one extremely dominant category. So it makes sense to recode every other value as 1 creating a binary variable of 'Reported no days with poor mental health' as 'MENTHLTH2'. The reasoning for doing this is that we are concerned that since relatively few people reported days with poor mental health leaving the granularity of number of days will lead to the model picking up a lot of noise, so creating a binary variable helps stem that risk.

Response Variable (Diabetes)

Let us recode this variable as a binary variable with Diabetics and Prediabetics coded as 1 and non-diabetics and Gestational diabetics coded as 0. Additionally, let us drop non-responses entirely. If there is no response for our explained variable we should drop those observations from our model.

Building the Model

Let us build a model to determine the risk factors for Diabetes and Pre-Diabetes.

Let's build the model using a logistic regression framework. This method has several advantages:

- It is easily interpretable to non-technical stakeholders, which will allow the model and its results to be utilized in the field, unlike black box methods like neural networks • The output will be expressed as a probability of an individual being diabetic or prediabetic. This will allow stakeholders to understand the likelihood of an individual being prediabetic
- The model will return coefficients for each variable indicating whether that attribute is a risk factor for diabetes

Our data is highly asymmetric, meaning we have many fewer diabetic and prediabetic people than not diabetic. Thus we will iterate our model to optimize for the true positive rate (recall) minus the false positive rate

There are several logistic regression models to use let's use sklearn LogisticRegression

This model allows us to use a machine learning framework to ensure our model has predictive ability, thus allowing stakeholders and practitioners to use our model to evaluate the risk of diabetes in new patients. This will be achieved by splitting our data into test and training sets, using the training set to fit the model, and using the model to predict values in the test set.

Metrics to assess performance

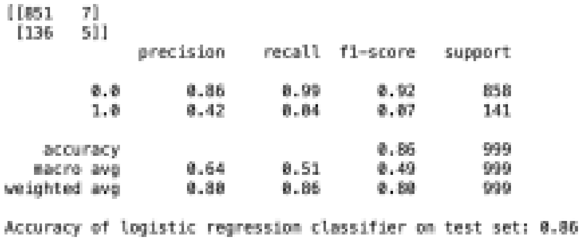
Predictive analytics - to access the predictive capability we will use the following metrics on the test set

- accuracy of the model on the test set ie number of correct predictions/total predictions
- recall (sensitivity or true positive rate): number of actual positive values the model detects/total actual positive values
- precision (specificity): number of actual positive values the model detects/total predicted positive values
- Comparing the performance on the test vs train sets: If the model performs much better on the training set than the test set model is overfitted, the model is picking up on noise in the training set

Determining risk factors

coefficient: magnitude/size of the coefficients will indicate the 'impact' of an attribute on the risk of diabetes

We will be converting our data to categorical variables as dummy binary variables. The reason for this is even our 'numeric' data is binned into groups thus we need to treat them as categories.



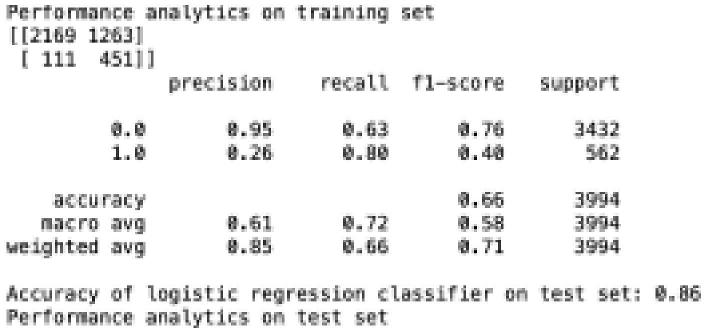
```
[[851   7]
 [136   5]]
              precision    recall  f1-score   support

         0.0       0.86      0.99      0.92       858
         1.0       0.42      0.04      0.07       141

    accuracy                           0.86       999
   macro avg       0.64      0.51      0.49       999
weighted avg       0.80      0.86      0.80       999

Accuracy of logistic regression classifier on test set: 0.86
```

Interpreting the confusion matrix and classification report

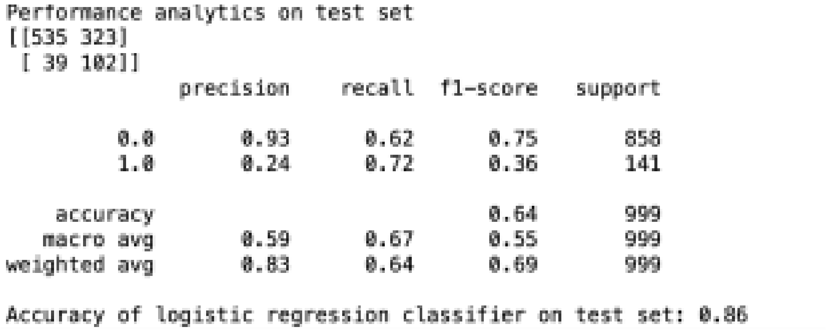
From our analysis, we can show that our current classification is insufficient in several ways. An accuracy of 86% seems good on the surface. However, our recall is only 4% which is very low. This indicates the model is doing a woeful job of detecting individuals with diabetes or prediabetes. This is especially problematic in a medical context where it is more important to identify people with risk factors than not identify people without risk factors.

I believe that this is primarily a function of the asymmetric nature of our data, meaning we have many fewer diabetic and prediabetic people than not diabetic. To remedy this let's optimize our model for True Positive Rate (Sensitivity) - False Positive Rate by changing the classification threshold, the probability at which the model will classify an individual as Diabetic or Prediabetic, instead of the default threshold (0.5)

threshold: 0.12424448671549021, optimal true positive rate - false positive rate 0.434484110195854



```
Performance analytics on training set
[[2169 1263]
 [ 111  451]]
              precision    recall  f1-score   support

         0.0       0.95      0.63      0.76      3432
         1.0       0.26      0.80      0.40       562

    accuracy                           0.66      3994
   macro avg       0.61      0.72      0.58      3994
weighted avg       0.85      0.66      0.71      3994

Accuracy of logistic regression classifier on test set: 0.86
Performance analytics on test set
```



```
Performance analytics on test set
[[535 323]
 [ 39 102]]
              precision    recall  f1-score   support

         0.0       0.93      0.62      0.75       858
         1.0       0.24      0.72      0.36       141

    accuracy                           0.64       999
   macro avg       0.59      0.67      0.55       999
weighted avg       0.83      0.64      0.69       999

Accuracy of logistic regression classifier on test set: 0.86
```

After we optimize our regression for true positive rate - false positive rate, our model performs much better. Our recall (sensitivity) is much higher at 72% meaning our model detects diabetes and prediabetes in 72% of individuals who have it. Notably, our precision is only 24%. In a medical context, this is okay because of the cost of a false positive is less than a false negative.

Additionally, the accuracy of the model is still 86%, so the optimized model does not have less accuracy relative to the base model

Finally, the model performs reasonably well on the test set relative to the training set. The overall accuracy of the models is almost equivalent (86% and 86%) the precision/specificity of the model is also relatively similar at (24% vs 26%). However, sensitivity on the test set is less than that on the training set (72% vs 80%). This is somewhat problematic as sensitivity is perhaps the most important performance metric. We can say that our model is slightly overfitted, but still has

strong predictive capacity.

Interpreting the model and finding risk factors for diabetes

From our coefficients, we can discern the 'impact' of each variable on the model. The greater the coefficient magnitude (size), the more impactful the model suggests the variable is. Positive coefficients indicate a greater risk of diabetes and negative coefficients indicate less risk of diabetes.

Income: The model suggests higher income people are at decreased risk of diabetes

Mental Health: The model suggests those who reported days with poor mental health are less likely to have diabetes. I believe this could be an instance of confounding variables

Race: The coefficients suggest Black/African Americans and Native Americans/Alaskans are at a greater risk for diabetes. Also note, that those who refused seem to be at a greater risk of diabetes

Asthma: Asthma is also a risk factor for diabetes

Smoking: The model suggests current smokers were at less risk for diabetes, which runs counter to what we expect. However, former smokers are at a greater risk for diabetes.

Marital Status: Marital status showed us an interesting relationship. Married and Widowed people showed the greatest risk, followed by divorced or separated people, then never married, and people in unmarried relationships. I suspect this is primarily a function of age. If we iterated the model, I would examine the relationship between marital status and age as well as age and diabetes.

General Health: The coefficients suggest those who reported better general health were at decreased risk of diabetes

Sex: The coefficients suggest males are at slightly greater risk for diabetes

What the model tells us, what limits, and what steps to improve the model in the future

Because we tested our model on a testing set we know the model has good predictive ability. The best use case would be for a practitioner to use our model to determine the probability of a patient having diabetes. Additionally, the model did discern potential risk factors that stakeholders could apply in the field.

However, the limitation of our model is the rigorousness with which we know the identified risk factors are actual risk factors for diabetes. I believe there is evidence that the model is confounding risk factors. To have more clarity on the risk factors, the next steps would be to run a significance test (Wald test and or t-test) on the variables to determine if the risk factors are statistically significant.

Other steps with a greater time scope would be to iterate the model with new combinations of variables including new variables, removing statistically insignificant variables. I'm specifically interested in adding age and removing marital status. This process would lead to a better-performing model and hopefully, one that is more explanatory.