



Predicting Diabetes from CDC Survey Data

Dhruv Ragunathan – Director of Data Solutions



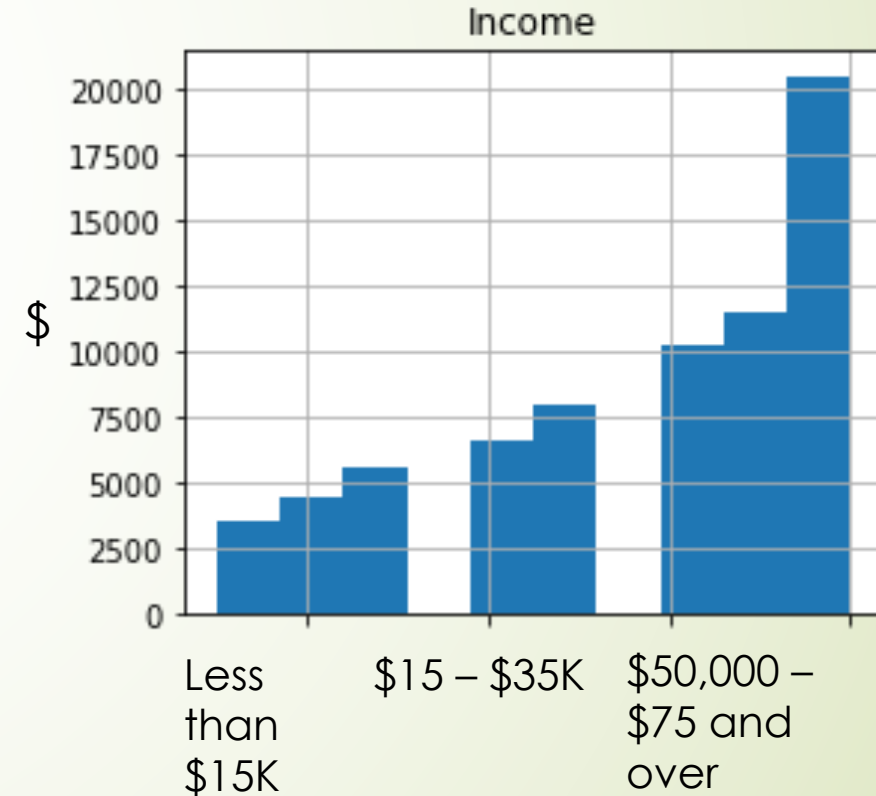
Business Objectives



- We have been tasked by the CDC to use the data collected from the Behavioral Risk Factor Surveillance System (BRFSS) to create models that predict diabetes.
- Create a model to tell respondents if they are at risk after the survey.
- Plan to create an app where people can screen online for their diabetic risk.
- The model will be tuned to reduce 'false positives' since these will waste surveyors time and resources.

Data Overview

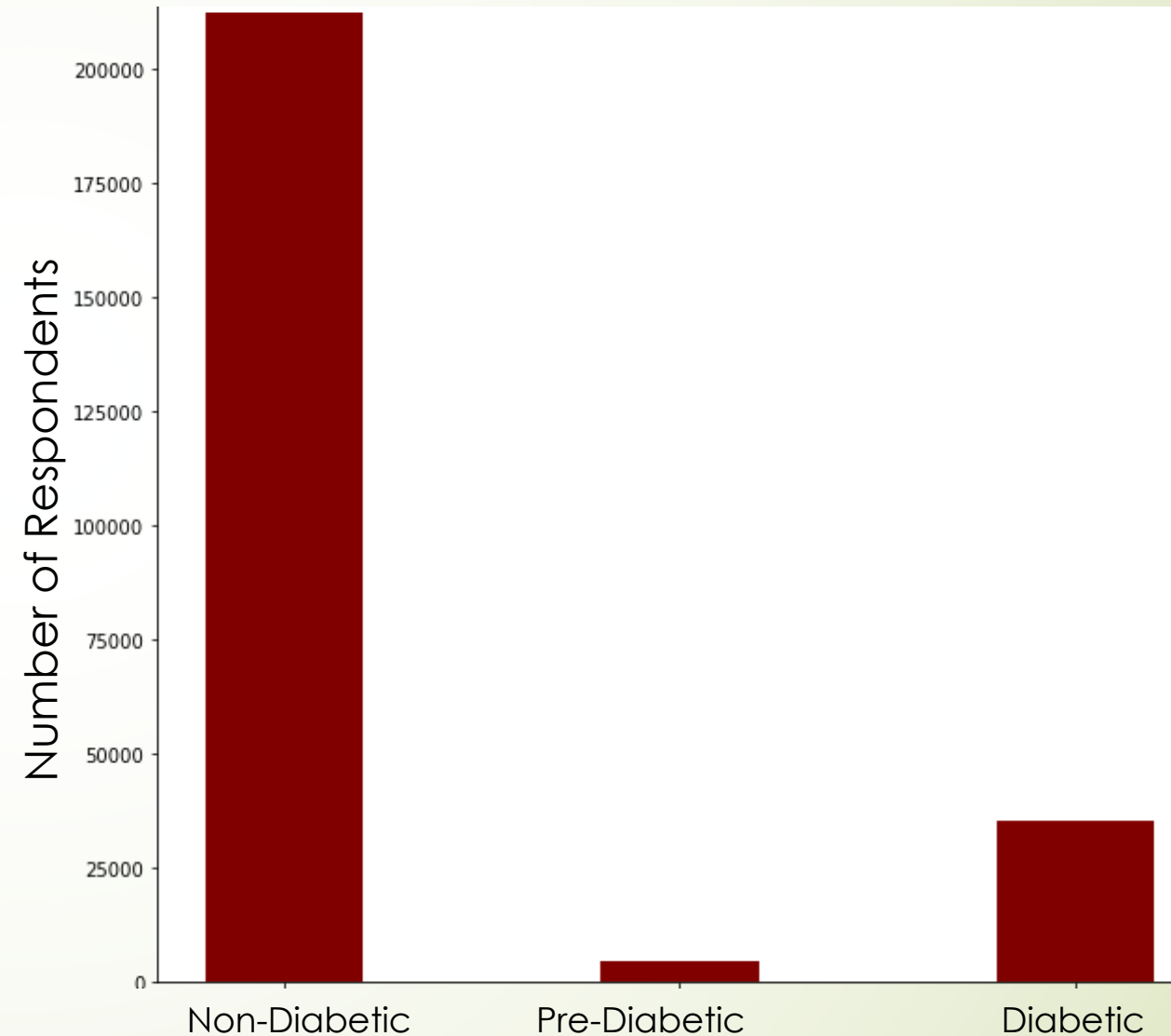
- The data was pulled from the 2015 BRFSS questionnaire.
- The data had around 440 thousand responses and 330 features.
- Some limitations off using this data to predict diabetes are:
 - Some survey respondents may not answer truthfully.
 - Some survey respondents may not be aware that they are pre-diabetic/diabetic.
 - Data seemed skewed towards higher – income respondents.



Data Preparation

- Created a set of columns we would model.
- Not all columns were relevant to diabetes prediction.
- Removed rows with missing values.
- Removed Rows that had responses such as "Refused", "Not sure", etc.
- Significantly more respondents are non-diabetic than diabetic.

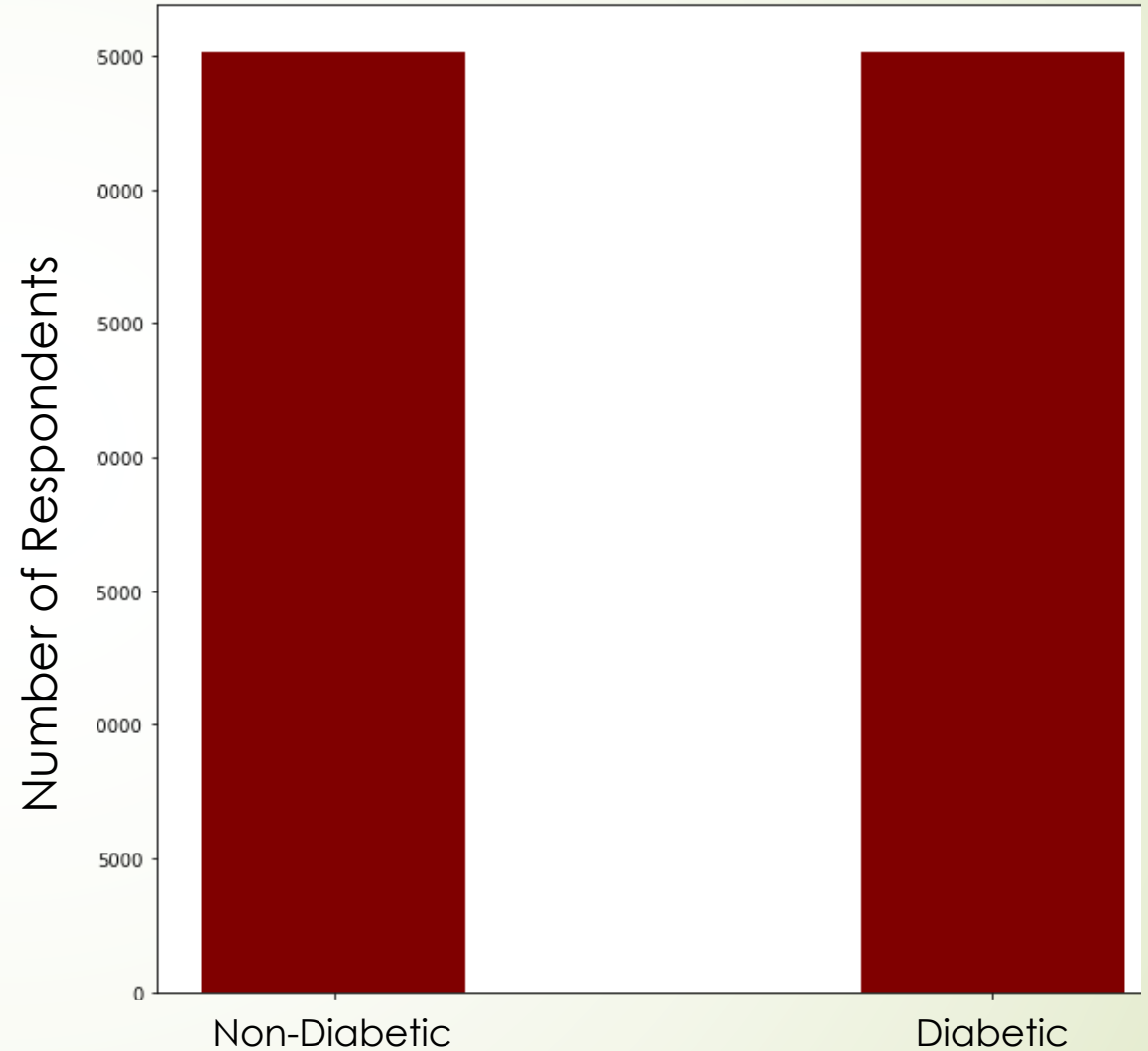
Respondents in Diabetic Categories



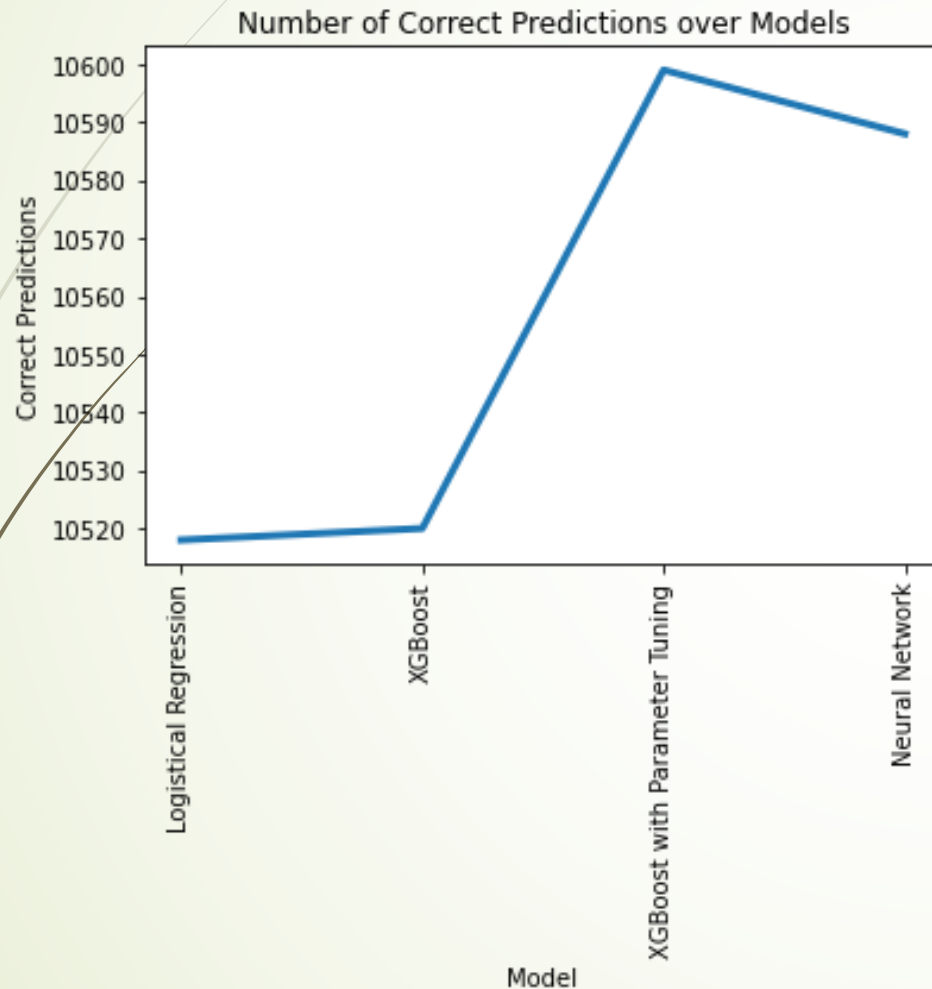
Data Preparation Cont.

- Dataset contained approximately 30,000 rows in each category.
- Consolidated pre-diabetic and diabetic features.
- The consolidation meets the business objectives to screen for respondents for diabetic risk.

Number of Respondents in Non-Diabetic, Diabetic, Prediabetic Categories



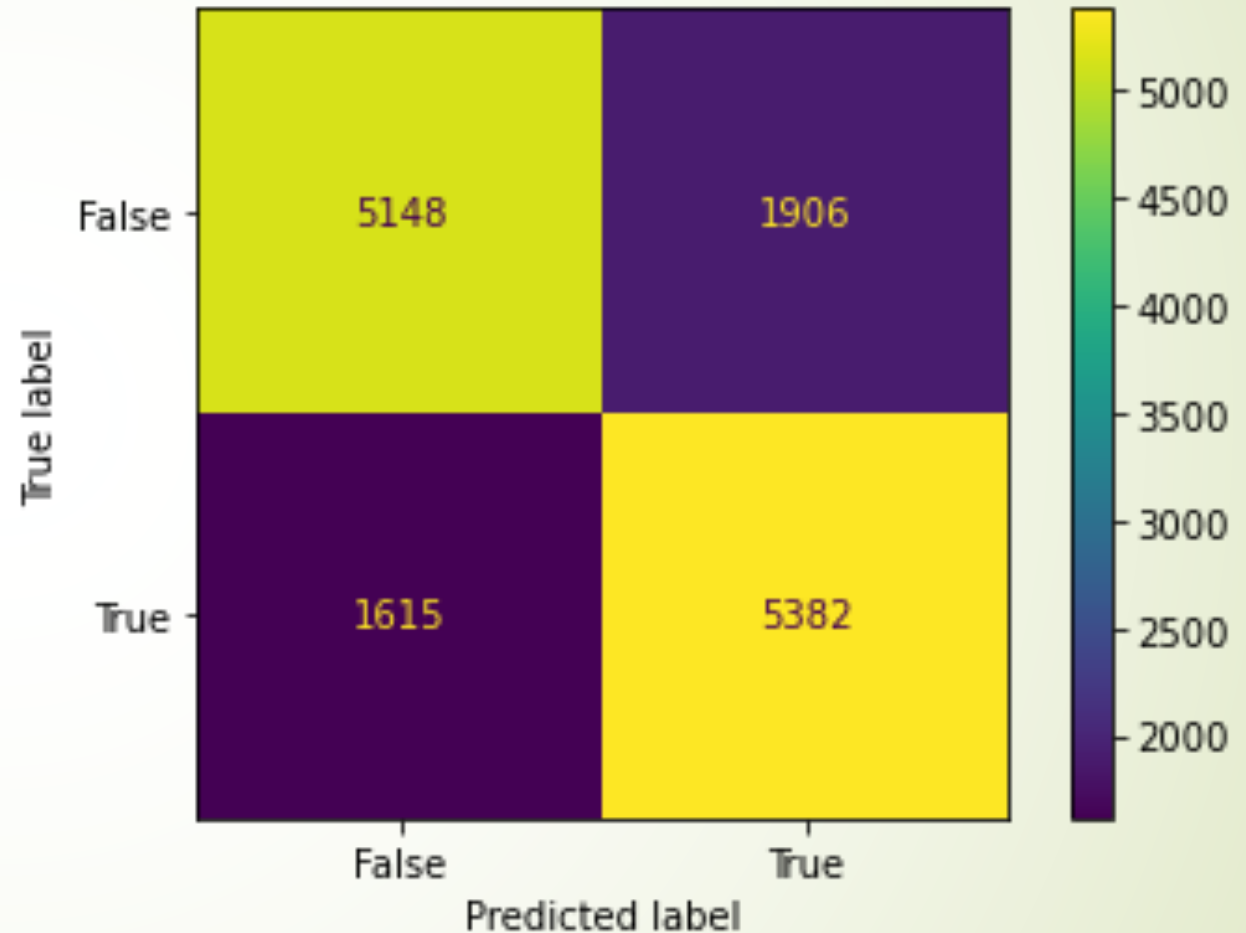
Model Results



	Accuracy	Precision	Runtime (s)
Logistical Regression	0.749	0.738	<1
XGB Tuned	0.753	0.732	38
Neural Network	0.760	0.744	48
SVC	0.753	0.728	350
Random Forest	0.749	0.724	7

Model Evaluation

- Logistical Regression (LR) was chosen as the model.
- The most accurate model was only 0.1% more accurate than it. Yet took significantly longer to run.
- Recommend using this model:
 - Accuracy and precision are close to other models
 - Easier to create
 - Significantly faster to run.





Recommendations



- The CDC should use the logistical regression model in their application.
- Consider a strategy around educating people to take their blood pressure on a regular basis since it was one of the top features.
- Providers who see people with high cholesterol should also screen for diabetes since high cholesterol was another top feature.
- Continue advocating for policy/strategies that aim to improve the general health and fitness of Americans. Low health was the most correlated feature with diabetes.



Future Work



- Evaluate previous BRFSS data sets. Measure the rate of diabetes and other chronic conditions to find their trends across the country.
- Leverage the model in a diabetic risk assessment online application.
- Add more features too the model such as race, sodium intake, etc.



Contact Information



- **Linkedin:** <https://www.linkedin.com/in/dhruv-ragunathan-908993b1/>
- **Github:** <https://github.com/dragunat2016>