Predicting Mental Health From Demographics and Lifestyle

Applying AI Algorithms to Identify At-Risk Individuals

Ali Black (ablack9), Kristen Anderson (ander50n)

Introduction

We are using demographic information and self reported mental health information to predict people who may be at risk of suicide or self-harm. In the training dataset, the prediction of "at risk" is determined by the individual's response to the question: Over the last 2 weeks, how often have you been bothered by the following problems: thoughts that you would be better off dead or hurting yourself in some way?

- Input: demographic and health information for 5924 individuals

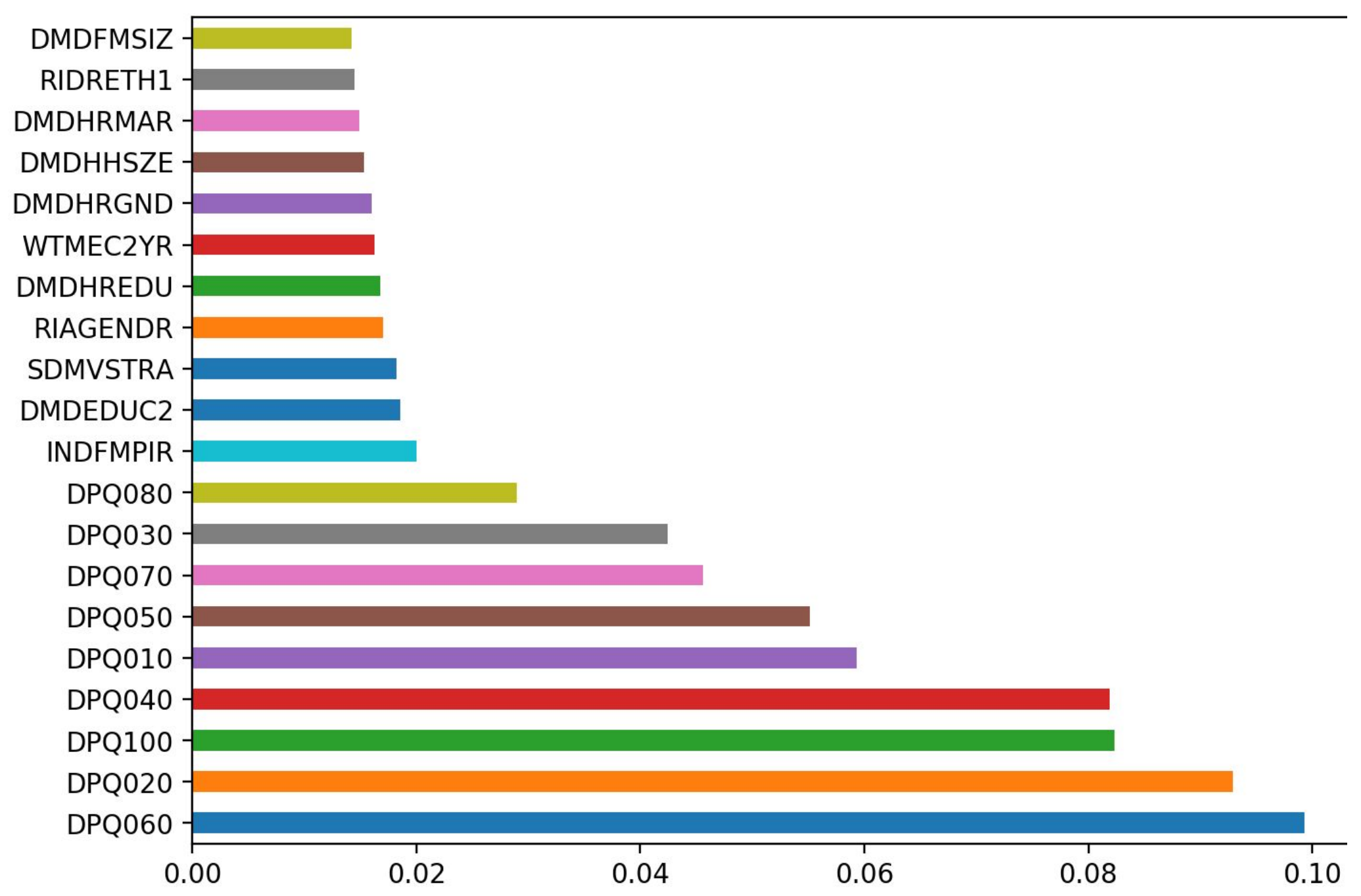
Data

The data comes from large set of health-care information that includes metrics of diet, demographics, physical health (both lab work and data collected through a professional physical exam), medications, and a comprehensive healthcare questionnaire. It is a combination of objective medical records and subjective self-reported answers.

Data Augmentation

Version	F1-Score 0 Prediction	F1-Score 1 Prediction	Accuracy
Logistic Regression	0.98	0.09	97%
SVM	0.98	0.09	96%
Logistic Regression with Data Augmentation	0.90	0.73	86%
SVM with Data Augmentation	0.88	0.60	80%

Feature Analysis



- critical features: mental health survey responses, poverty level, education, gender, marital status
- features removed: language factors in interview, other logistical questions, pregnancy status

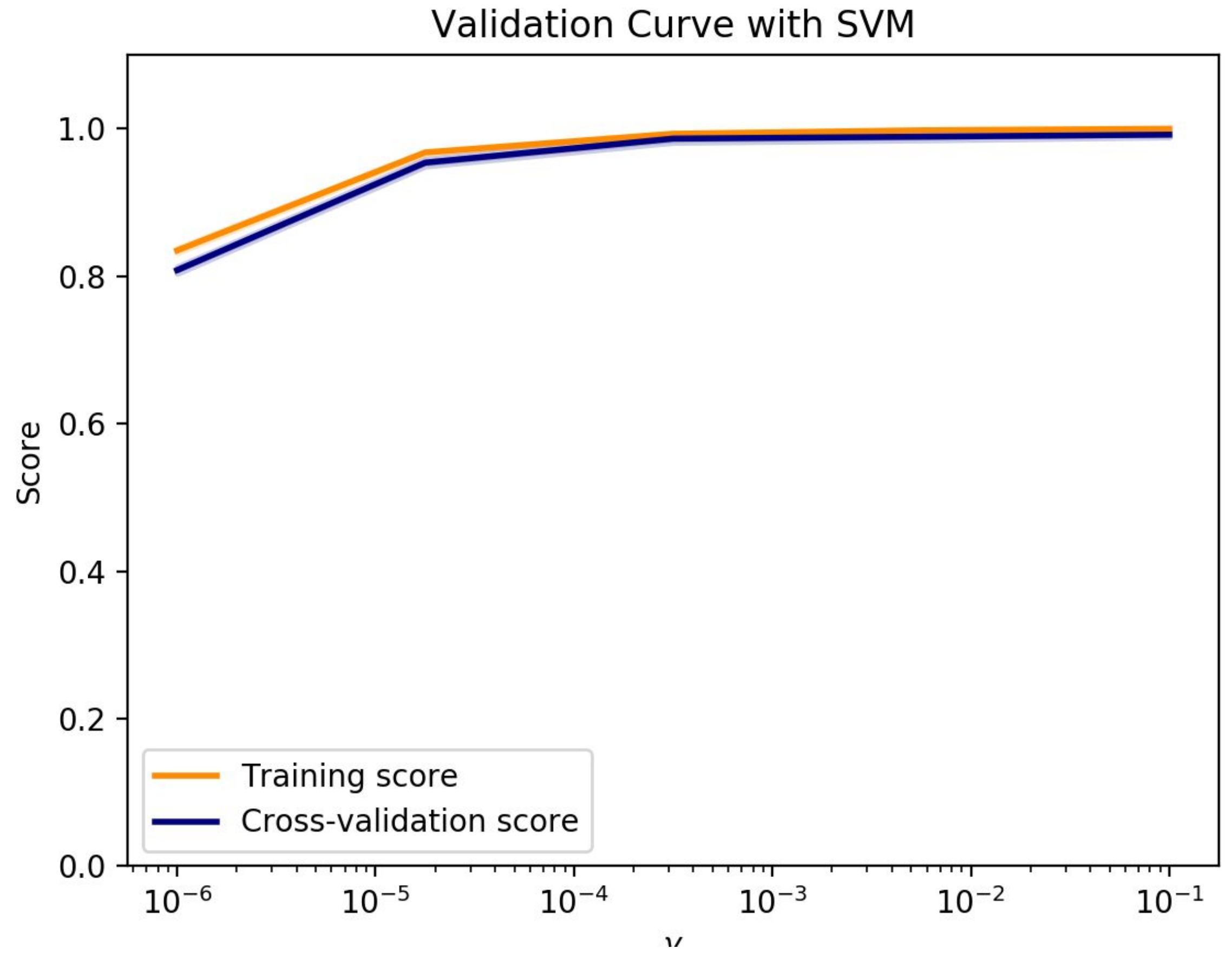
Implementation

We have implemented logistic regression and several variations of our SVM (Support Vector Machine) model to predict mental health status. SVM is used to create a binary classification, by bisecting a dataset with a line that is maximally distant from the points of the dataset. Our implementation of SVM relies on a linear division of the space.

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(\vec{w}\cdot\vec{x}_i - b)\right)\right] + \lambda\|\vec{w}\|^2$$

Hyperparameter Tuning

- experimented with polynomial, linear, gaussian, and sigmoid kernels
- optimal hyperparameters: {'C':1, 'gamma':0.1, kernel: 'gaussian'}
- as seen below, minimal overfitting using these parameters



Results

Version	F1-Score 0 Prediction	F1-Score 1 Prediction	Accuracy
Logistic Regression with D.A.	0.90	0.73	86%
Linear SVM with D.A.	0.88	0.60	80%
Gaussian SVM with D.A.	0.99	0.98	99%

Discussion

To more rigorous test our best model, Gaussian kernel SVM with data augmentation and feature extraction, we tested it on different demographic divisions.

Gender

- Men - 98% accuracy
- Women - 100% accuracy

Race

- Hispanic - 98% accuracy
- White - 99% accuracy
- Black - 99% accuracy
- Asian - 100% accuracy
- Other and Mixed - 86% accuracy

Our model performs well on this specific dataset. Given the specifications of the data, the model has little bias and classifies "at risk" individuals of different genders and races with the same accuracy. Due to the unique parameters of the dataset, our model is not generalizable to the real world, but it serves as a good model from which an institution could base an algorithm for flagging at risk individuals.

References

- https://towardsdatascience.com/support-vector-machine-python-example-d67d9b63f1c8
- https://www.quora.com/What-is-the-difference-between-Linear-SVMs-and-Logistic-Regression
- https://en.wikipedia.org/wiki/Support-vector_machine
- https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop.html
- https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey#questionnaire.csv
- https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e