GPOSPN (*GLOBAL PREVALENCE OF SCHIZOPHRENIA PREDICTIVE NETWORK*)

Jack Hudson

Introduction

This project consists of an interactable platform that uses various machine learning algorithms and a neural network to make predictions on the prevalence of schizophrenia in any country with available data. My partner, Zander Chearavanont (Email: *zchearavanont@exeter.edu*), and I set out to find which model would best predict the proportion of schizophrenia in the given country, and used that to make future predictions.

Data Set: https://www.kaggle.com/datasets/valchovalev/shareofpopulationwithschizophrenia

Gradient Boosting

Unlike simplistic single-model algorithms like linear or logistic regression, gradient boosting is an ensemble method that first uses sample groups to create a model and then deduces the patterns in their residuals or cost (cost can be thought of as how far off a prediction is to the actual value, the lower the better) to reduce them. It then repeatedly creates new models based on the residuals until the difference between the targeted and predicted values is very low or nonexistent. Then it takes the sum of all the created models which takes the initial predictive model and remodels it based on the derived errors. Gradient boosting proved to be decently accurate, having a consistent root mean squared error of less than 0.01.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

Exponential Smoothing

Proving to be the most accurate algorithm thus far, exponential smoothing is specifically designed to forecast time series data (data that changes over time). It creates predictions, weighted sums, using exponentially decreasing weight for past observations. The (????(alpha))

parameter which is the smoothing factor/smoothing coefficient (0-1) dictates the rate at which the influence of the observations made during prior steps decay exponentially. The GPOSPN used alpha values of 0.4, 0.6, 0.8, and an alpha that was deduced by the stats model module. While all the predictions were quite accurate, the alpha value deduced by the stats model unsurprisingly turned out to be the most accurate.

$$\begin{aligned} \widehat{y}_{t+1} &= \ell_t && \text{(forecast equation)} \\ \ell_t &= \alpha y_t + (1-\alpha)\ell_{t-1} && \text{(smoothing equation)} \end{aligned}$$

for an initial value ℓ_0 and $0 \leq \alpha \leq 1$.

LSTM (Long-Short Term Memory)

We were inspired to use an LSTM model by researchers in a paper named Detection of dementia on voice recordings using deep learning (https://link.springer.com/content/pdf/10.1186/s13195-021-00888-3.pdf) to use an LSTM model. An LSTM is a type of RNN (recurrent neural network), which typically deal with sequenced datasets (where each data point depends on the previous data point). Since a schizophrenia proportion is potentially dependent on previous years, RNNs are perfect for our time-series predictions. Where they fall short, however, is in their short-term memory. To put this into perspective, in a program used to predict the next word in a sentence, a traditional RNN would only account for the words stored in the neurons of memory cells (which hold the word) nearby and not the entire sentence. This is solved by using the long-term memory in LSTM models as it stores keywords that aren't necessarily near the word before the prediction.

$$\begin{aligned} f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\ h_t &= o_t \circ \sigma_h(c_t) \end{aligned}$$

Prevalence of Schizophrenia & Conclusion

Schizophrenia is a mental disorder that consistently facilitates a state of psychosis in those who have it. It can lead to a variety of symptoms such as hallucinations, delusions, paranoia, disorganized thinking, social withdrawal, decreased emotional expression, and apathy. As long as there is good quality data for training, projects similar to the GPOSPN can help nations or areas predict whether or not schizophrenia, or similar illnesses, will be a prevalent issue. The models presented in the GPOSPN can also be used to test whether it's even possible to predict certain mental disorders. Schizophrenia is uncommon and under-researched with no definite biomarker, yet when we train our models with early years, i.e. 1990-2010, and compare predictions with later years, i.e. 2011-2017, results are remarkably similar with certain models.

Further Work

Our current display (the graphs) merely displays the accuracy of our models (which were previously tested with the regular split of training, validation, and testing sets before we allowed them to use the entire dataset to make predictions) and the "prediction" for 2018 (our dataset stops at 2017) though the GPOSPN can predict the future indefinitely. We hope to implement a website with a world heat map that displays the prevalence in every country with a scroll bar that can be moved back and forth through time to see the prevalence change in each year. We also hope to implement buttons that show alternative heatmaps for males and females (there is no data on the prevalence amongst nonbinary individuals in the dataset).

Directions

To access the project on your end, follow these steps:

1. Open google drive and sign in to your google account.
2. Go to this page: Schizo_df.csv
3. Click the download button.
4. After *schizo_df.csv* appears in the bottom left of the screen, navigate to your google drive, then click and drag the box under *files*.

5. Go to this page: https://colab.research.google.com/drive/1RtdZT1By5rbVOU-Ur5-iE-nj2xokZ-xg?usp=sharing
6. Click on *Runtime* in the top right and select *Run all*.
7. Follow the given prompts.
8. Type in any country you can think of, using capitalization when necessary.
9. Repeat step 6 to input a new country.

If you have any additional questions about the directions or any part of the project as a whole, contact jwhudson@umass.edu