Predicting Parkinson’s Disease based on Vocal Recordings

Vilo Avila - Data Scientist

Problem Statement

Parkinson's disease is a progressive disorder that affects the nervous system and the parts of the body controlled by the nerves.

My goal for this project is to find a model that can predict Parkinson's disease in a patient to the highest degree of accuracy based on features derived from vocal recordings, which could then be possibly used as an additional screening tool in a hospital or examination room setting.

PARKINSON'S DISEASE

A disease that affects nerve cells in the brain and causes tremors, poor coordination, and problems walking and moving

CAUSES & RISK FACTORS

Both sexes & all races are affected

Parkinson's commonly develops after age 50

Scientists have identified abnormal genes that may lead to parkinson's in some people, but there is no solid proof to show it is always inherited

Men are more likely to develop parkinson's disease because they're more likely to experience head injury or exposure to toxins

SYMPTOMS OF PARKINSON'S

- Slow blinking
- No facial expression
- Drooling
- Difficulty swallowing

- Shaking, tremors
- Loss of small or fine hand movements

- Memory loss, dementia
- Anxiety, depression
- Hallucinations

- Stooped posture
- Aches and pains
- Constipation
- Problems with balance or walking

The Data

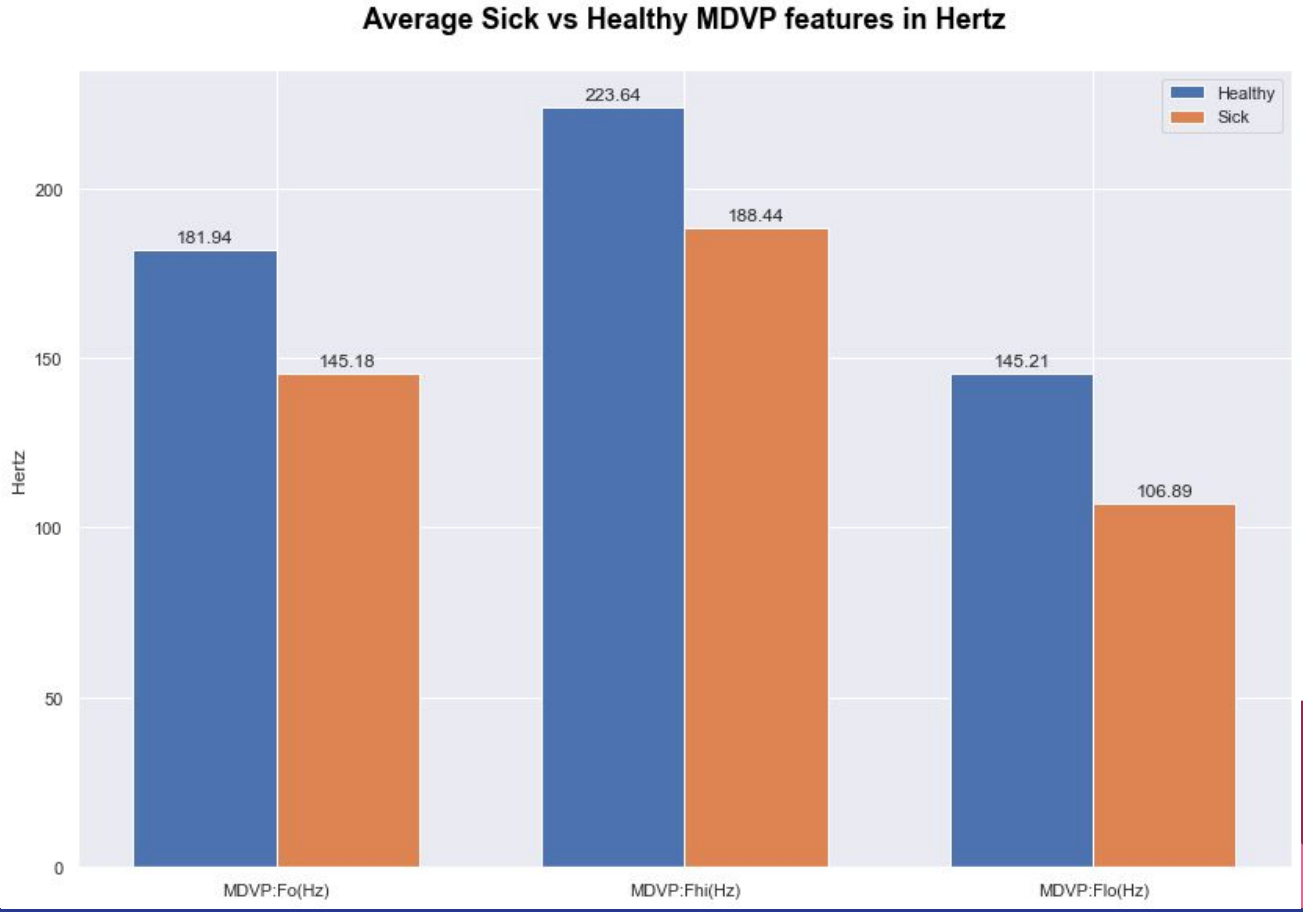
- **This dataset is composed of a range of voice measurements from 31 people, 23 with Parkinson's disease (PD)**
- **Each column in the table is a particular voice measure, and each row corresponds to one of 195 voice recording from these individuals.**
- **The dataset was downloaded from UC Irvine dataset repository.**

Cleaning

- Data set consists of 195 rows and 24 columns
- The dataset was cleaned by checking for null values, data types, and duplicates
- Outliers were checked but I ultimately kept in the dataset.

Exploratory Data Analysis

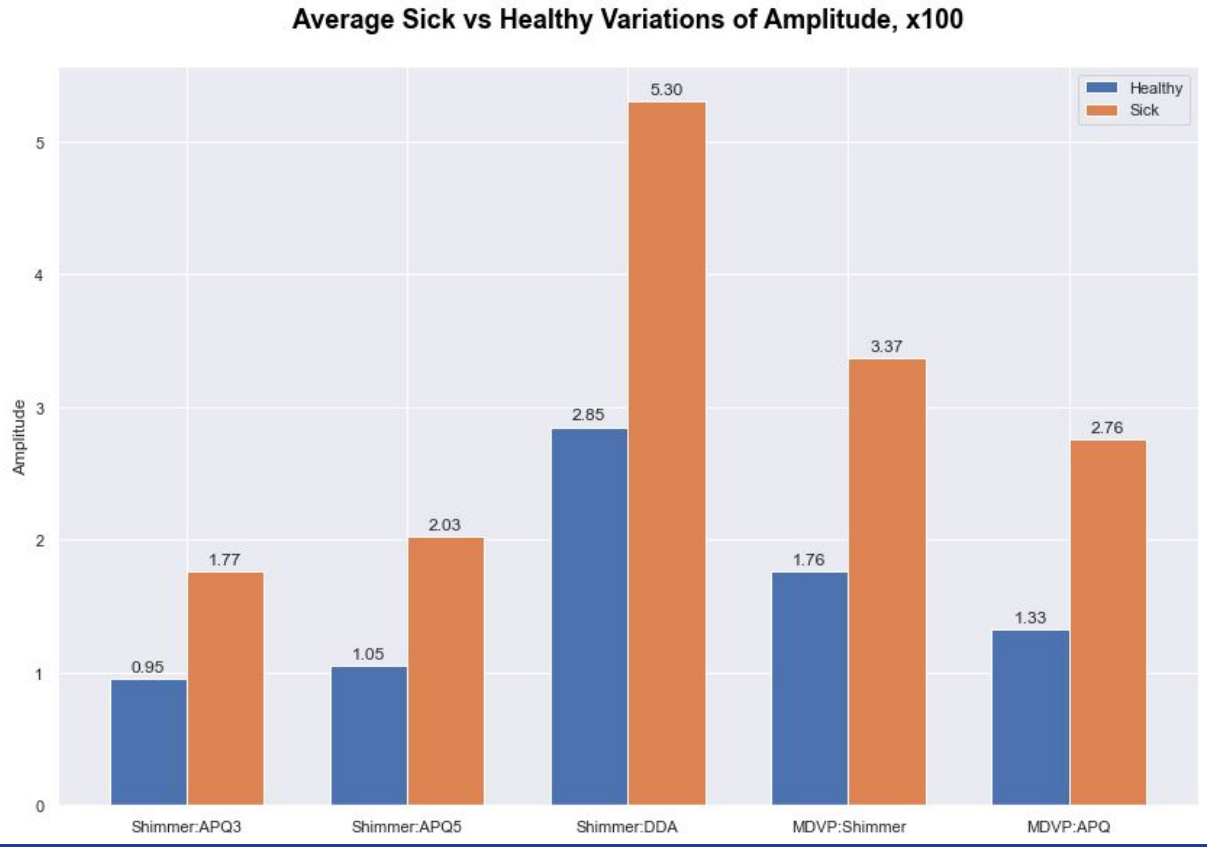
- **MDVP: Multidimensional Voice Program**
- **Fo - Average frequency**
- **Fhi - High Frequency**
- **Flo - Low frequency**



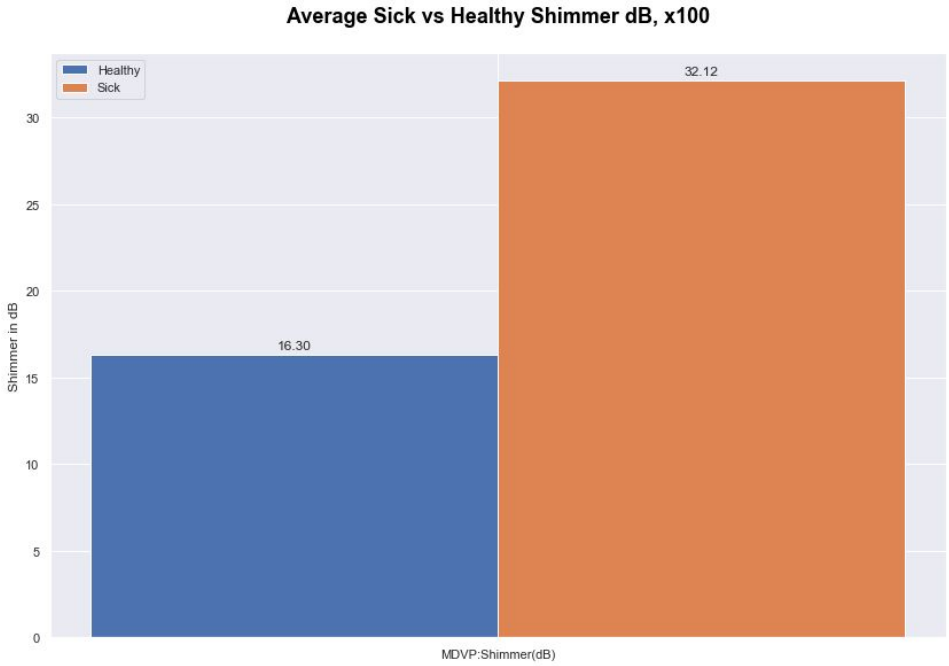
EDA continued

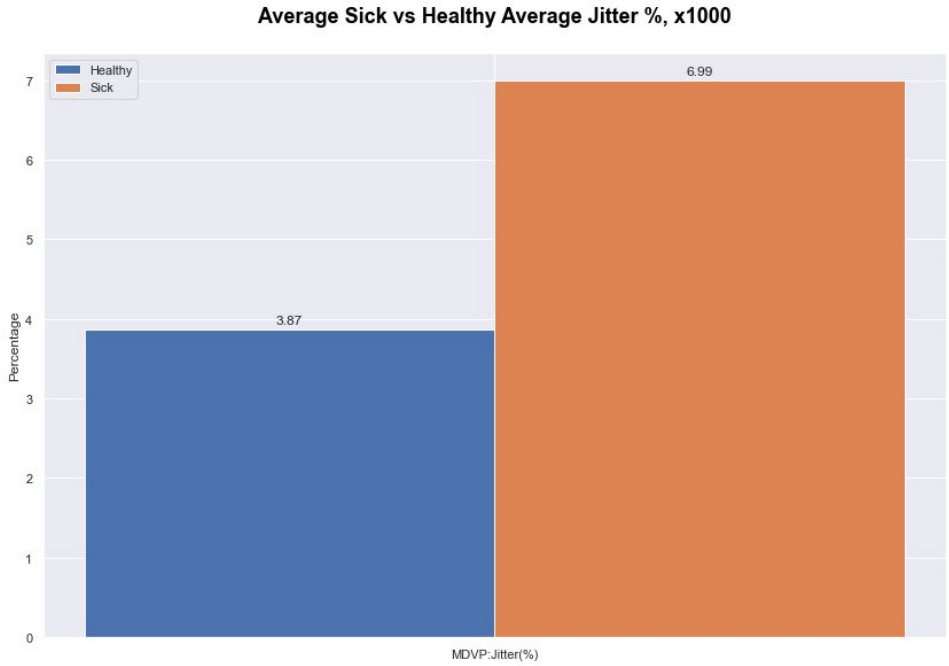
- MDVP: Multidimensional Voice Program
- Shimmer - measurement of changes in the amplitude of the speech wave.
- APQ - Amplitude Perturbation Quotient
- DDA - Average absolute difference between consecutive periods

Summary: Several Measures of variation in amplitude



EDA continued





Machine Learning Models

- Data was preprocessed for use in modeling, this helps our model perform better overall and improve predictions

- Random Forest
  - Accuracy: 93.33%
- Support Vector Machine
  - Accuracy: 97.95%
- Gradient Boosting
  - Accuracy: 98.46%

Conclusions and Recommendations

- Given the outcome of our models, it would seem Parkinson's disease can be successfully predicted using recordings of a patient's voice. I feel this would be an excellent tool for a doctor to use while in early stages of diagnosing if a patient may have Parkinson's disease. I believe that with some further work I would be able to build an algorithm that could take in raw vocal data and feed it through a pipeline where everything gets processed and sent through to the final production model.
- As far as which model to choose - the Gradient Boosting had the highest accuracy out of all of our models with a score of 98.46%. We could use this for production.
- I recommend investing time and money in my work in order to complete these goals so we could have a non-invasive tool that could quickly and accurately determine if a patient may have Parkinson's disease. This could be used in a doctor's office or hospital setting where they could ask a the patient to say a predetermined phrase in a microphone and get a result in minutes.