Predict Liver Failure based on People's Demographics

Prashant H Krishnan

April 07, 2019

Introduction

The aim of this document is to explore the idea of being able to build a Machine Learning / Deep Learning model which can be used to predict the possibility of a liver failure in individuals given their demographics and health related information.

Business Problem

- The liver is the largest organ in the body of all vertebrates including humans which performs several important functions for the healthy functioning of the body.
- The failure of a liver is a life threatening condition that demands urgent medical care.
- However if detected early enough, supportive medical care can be given to treat the symptoms and to try and see if its effects can be reversed.

Data

- The Liver failure dataset from JPAC Center for Health Diagnosis and Control was downloaded from Kaggle.
- Data contains demographic and health information collected via surveys, direct interviews, examinations, and blood samples.
- This dataset consists of data from adults 20 years of age or older taken from 2008–2009 and 2014–2015 surveys.

Feature Set

- Age
- Gender
- Body Mass Index
- Waist
- Maximum Blood Pressure
- Minimum Blood Pressure
- Good Cholesterol
- Bad Cholesterol
- Total Cholesterol
- Dyslipidemia
- PVD
- Physical Activity
- Poor Vision
- Alcohol Consumption
- Hyper Tension
- Family Hyper Tension
- Diabetes
- Family Diabetes
- Hepatitis
- Family Hepatitis
- Chronic Fatigue

Solution

- Our model is an example of a Binary Classification Model
- 1 or 0 indicating “liver failure” or “no liver failure”
- Algorithms used for prediction model
  - Supervised Machine Learning
    - Gradient Boosted Trees
  - Deep Learning
    - Feed Forward Neural Network – Multi Layer Perceptron

Model Performance

- Gradient Boosted Trees (Supervised Machine Learning Model)

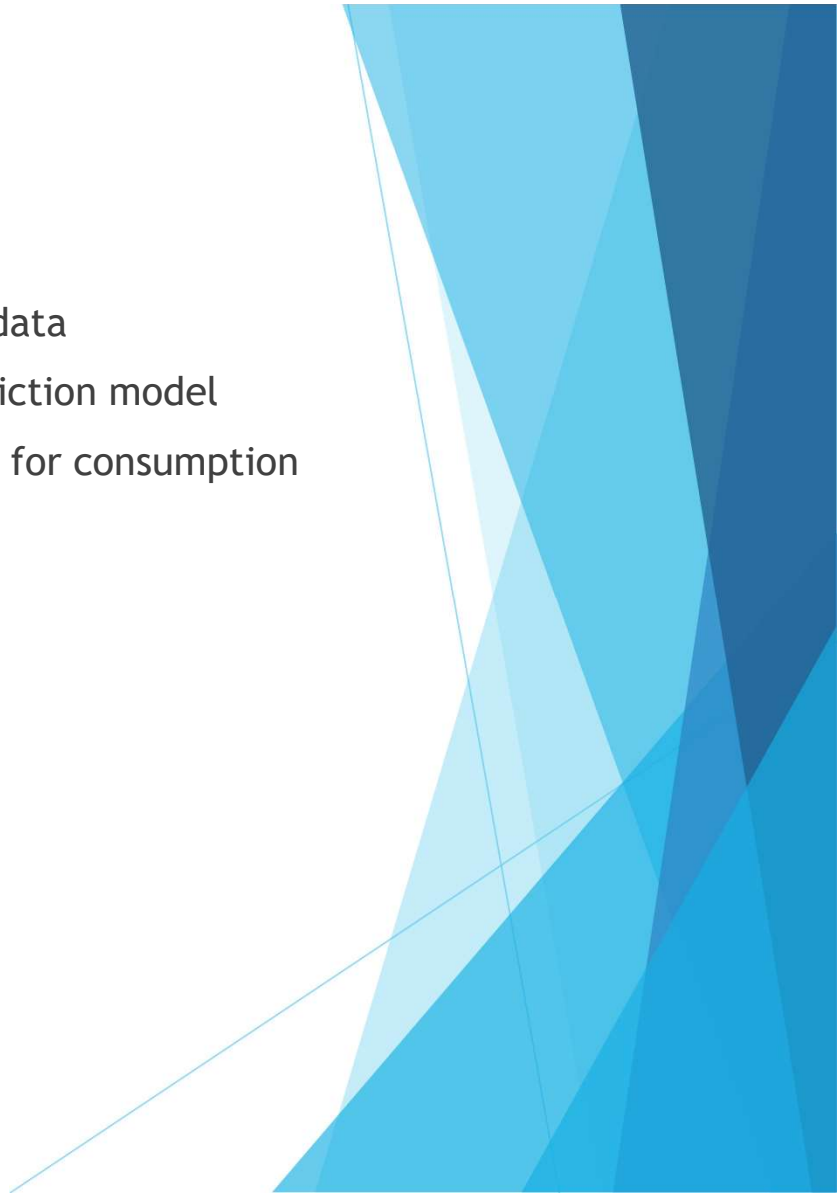
Training Accuracy	**96%**
Validation Accuracy	**92%**

- Multi Layer Perceptron (Deep Learning Model)

Training Accuracy	**94%**
Validation Accuracy	**93%**

Next Steps

- Continue to train and tune prediction models based on new data
- Look at streaming options to feed real time data to the prediction model
- Build a complete API based solution encapsulating the model for consumption by data products working in this space



Technical Details

Architectural Choices

- Simple Architecture comprising of the following components:

Component	Technology Choice
File Storage	IBM Cloud Object Storage
Data Repository	IBM Cloud Object Storage
Development Platform	IBM Watson Studio
	Apache Spark Framework (Spark 2.3)
	Jupyter Notebooks with Python 3.5
Data File Format	CSV

Data Quality Assessment

Data Quality Issue	Solution	Description
Missing Label Data	Filtering	Removed rows missing label data. This is our target variable and this value cannot be missing in our dataset being used for training and validation.
Missing Features Data	Imputing	Data for some of the feature columns were manually updated based on the value distribution.
Missing Features Data	Filtering	Columns for which data could not be calculated or inferred, the corresponding rows were removed from dataset.

Feature Engineering

- The following was done as part of data pre-processing and feature engineering

	Description
Encoding	Categorical features with strings values were converted to numbers.
One Hot Encoding	The categorical feature variables were converted to encoded vectors to make them Machine Learning friendly.
Vector Assembler	All the feature columns / vectors were merged into a single features vector.
Normalizing / Scaling	All the features were scaled to a value range of 0 to 1.

Model Algorithm & Performance Indicators

Model Algorithm

- Our Use case and dataset is an example of a Binary Classification model
- The following algorithms were selected to build our Binary Classification model
  - Supervised Machine Learning – Gradient Boosted Trees
  - Deep Learning – Feed Forward Neural Network (Multi Layer Perceptron)
- Both these algorithms are powerful algorithms for Binary Classification.
- These algorithms gave us better results in comparison to other classification models.

Performance Indicators

- Considering that ours is a classification model, we decided to use accuracy as the metric for capturing model performance.

Reference

- Project Report

https://github.com/phkrish/AdvancedDataScienceCapstone/blob/master/Advanced%20Data%20Science%20Capstone%20-%20Project%20Report.pdf

- Architectural Decisions Document

https://github.com/phkrish/AdvancedDataScienceCapstone/blob/master/Architectural%20Decision%20Document.docx

- Project Implementation (Jupyter Notebooks)

ETL - https://github.com/phkrish/AdvancedDataScienceCapstone/blob/master/Capstone_PredictLiverFailure.etl.ipynb

Feature Engineering –

https://github.com/phkrish/AdvancedDataScienceCapstone/blob/master/Capstone_PredictLiverFailure.feature_eng.ipynb

Model (Definition, Training and Evaluation)-

https://github.com/phkrish/AdvancedDataScienceCapstone/blob/master/Capstone_PredictLiverFailure.model_def_train_eval.ipynb

THANK YOU!