Machine Learning Engineer Capstone Proposal

Heart Disease Prediction: Saving lives using Machine Learning

by Nidhi Pansuriya

A) Domain Background: Overview

Heart disease is the leading cause of death for people of most racial and ethnic groups in the United States, including African American, American Indian, Alaska Native, Hispanic, and white men. For women from the Pacific Islands and Asian American, American Indian, Alaska Native, and Hispanic women, heart disease is second only to cancer.

- One person dies every 36 seconds just in the United States alone from cardiovascular disease.
- About 655,000 Americans die from heart disease each year—that's 1 in every 4 deaths.
- Heart disease costs the United States about $219 billion each year from 2014 to 2015. This includes the cost of healthcare services, medicines, and lost productivity due to death. Data Source: Heart Disease Facts | cdc.gov

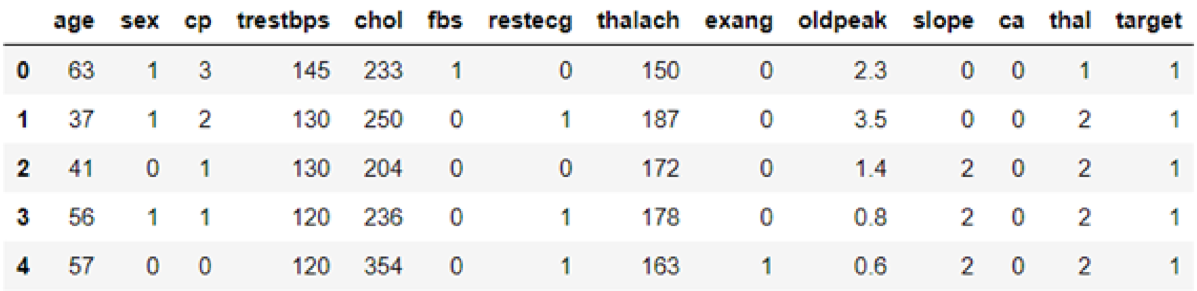
B) Problem Statements

- Complete analysis of Heart Disease UCI dataset both visually and statistically to obtain critical observations which can be used for inference.
- To predict whether a person has a heart disease or not based on the various biological and physical parameters of the body
- To make a model having high accuracy and precision and can predict the results with greater confidence.
- Make these predictions accessible to users and patients anywhere, anytime so that they can get complete picture of their Health

C) Datasets and Inputs

1. Collecting Data

The data used for training and testing is the Heart Disease UCI downloaded from Kaggle. The dataset consists of 303 individual data. There are 76 attributes in the dataset, however all published experiments refer to using a subset of 14 of them as far as machine learning is considered so. The "goal" field refers to the presence of heart disease in the patient.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

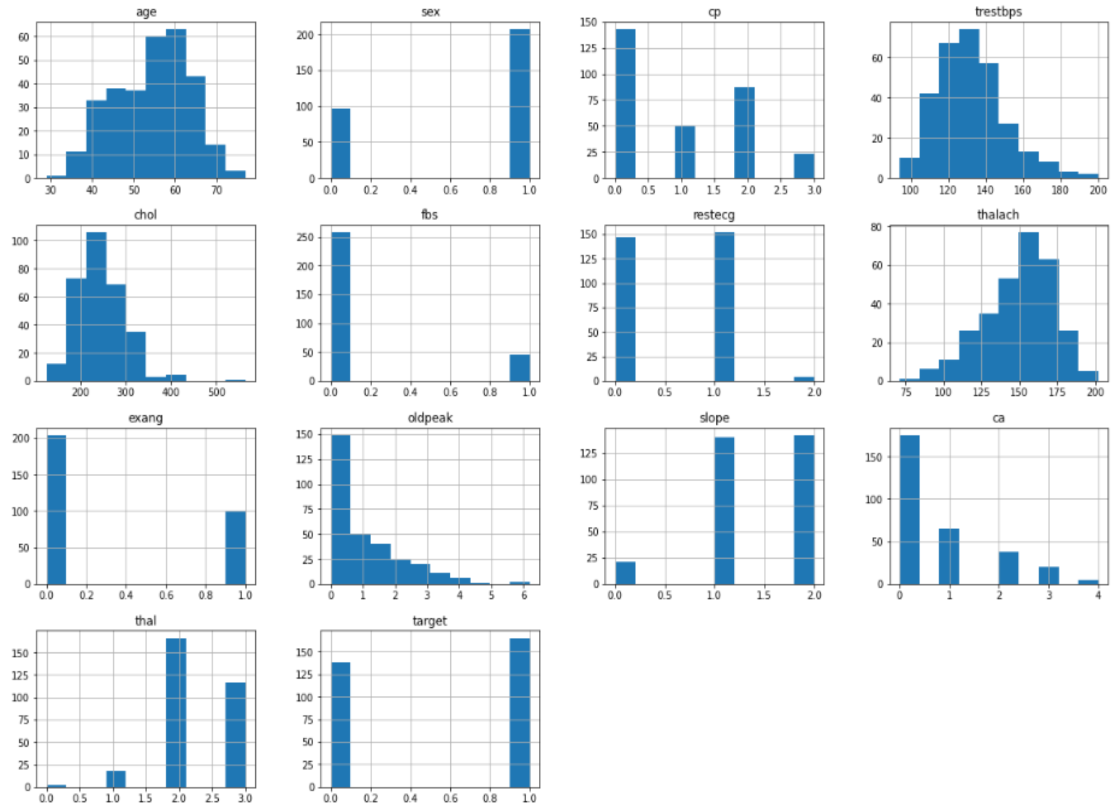
2. Exploratory Data Analysis

It's a clean, easy to understand set of data. However, the meaning of some of the column headers are not obvious. Here's what they mean,

- age: The person's age in years
- sex: The person's sex (1 = male, 0 = female)
- cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
- trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
- chol: The person's cholesterol measurement in mg/dl
- fbs: The person's fasting blood sugar (&gt; 120 mg/dl, 1 = true; 0 = false)
- restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
- thalach: The person's maximum heart rate achieved
- exang: Exercise induced angina (1 = yes; 0 = no)
- ldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
- slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
- ca: The number of major vessels (0-3)
- thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect)
- target: Heart disease (0 = no, 1 = yes)

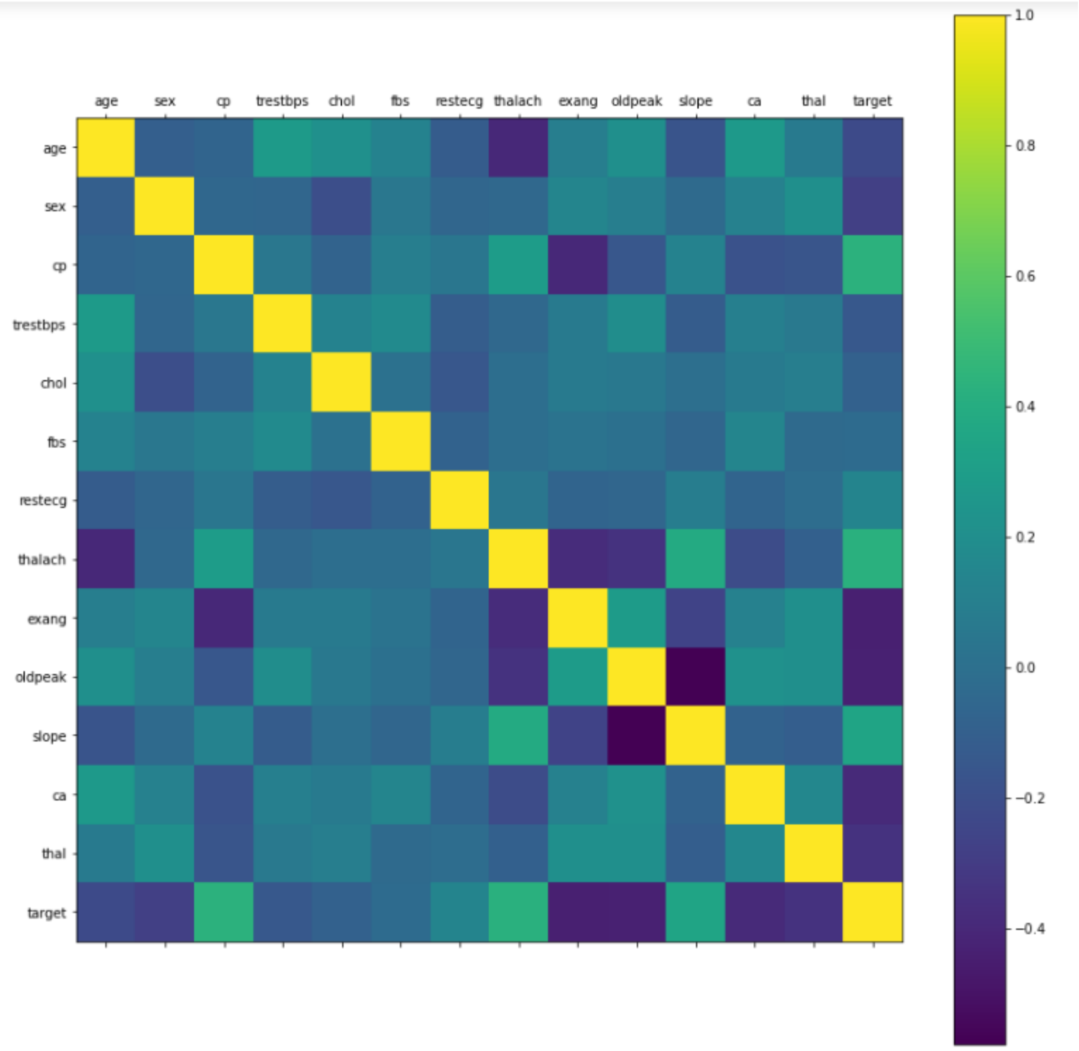
3. Data Visualization

Now let's see various visual representations of the data to understand more about various features.



4. Correlation Matrix

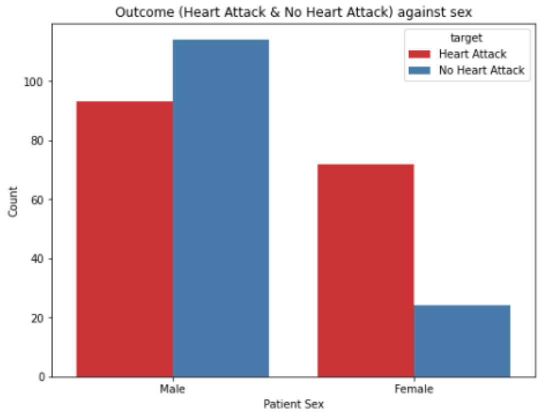
The best way to compare relationships between various features is to look at the correlation matrix between those features.



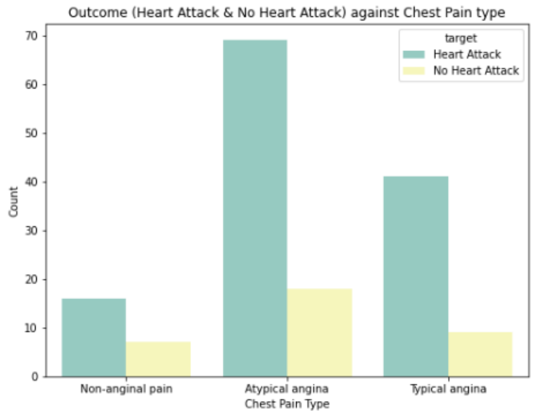
5. Inputs

Here our Model is trained to predict whether a person has a heart disease or not based on the following common features as input:

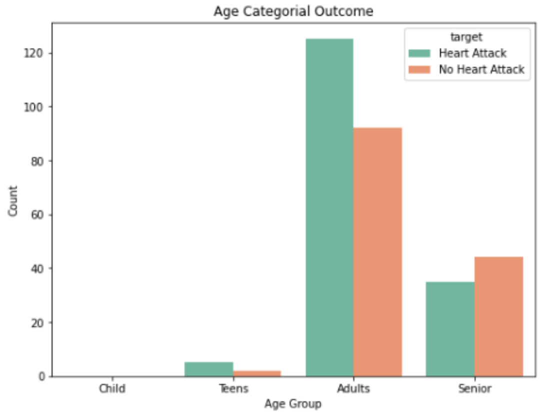
- age
- gender
- chest pain
- blood pressure
- cholesterol level
- blood sugar
- max heart rate



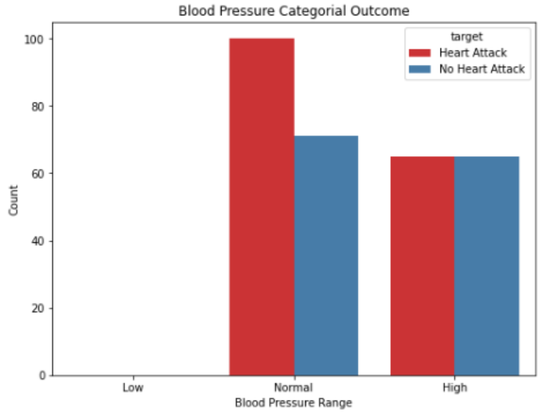
Outcome (Heart Attack & No Heart Attack) against sex



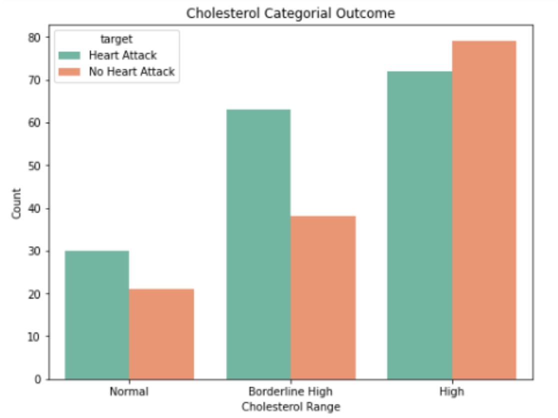
Outcome (Heart Attack & No Heart Attack) against Chest Pain type



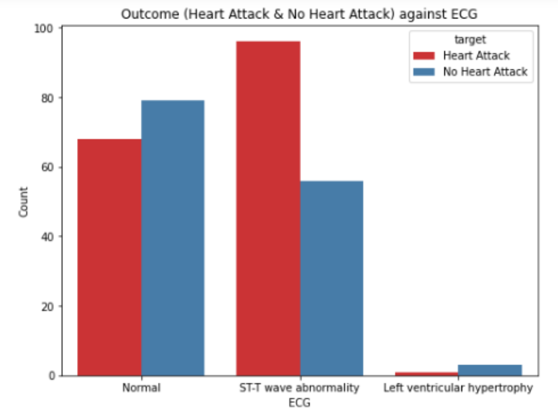
Age Categorial Outcome



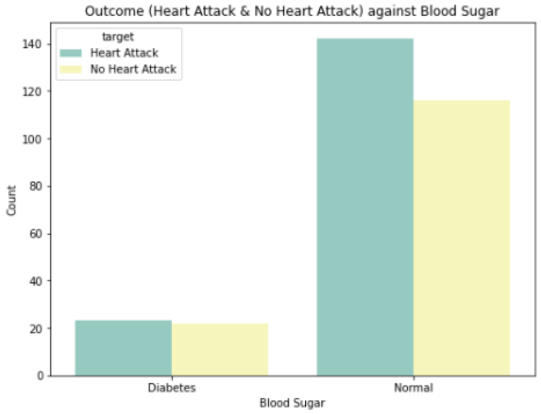
Blood Pressure Categorial Outcome



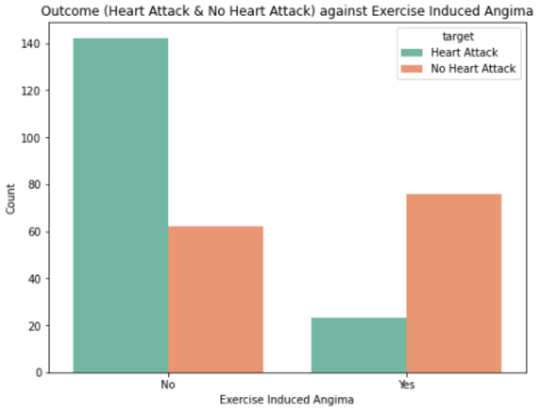
Cholesterol Categorial Outcome



Outcome (Heart Attack & No Heart Attack) against ECG



Outcome (Heart Attack & No Heart Attack) against Blood Sugar

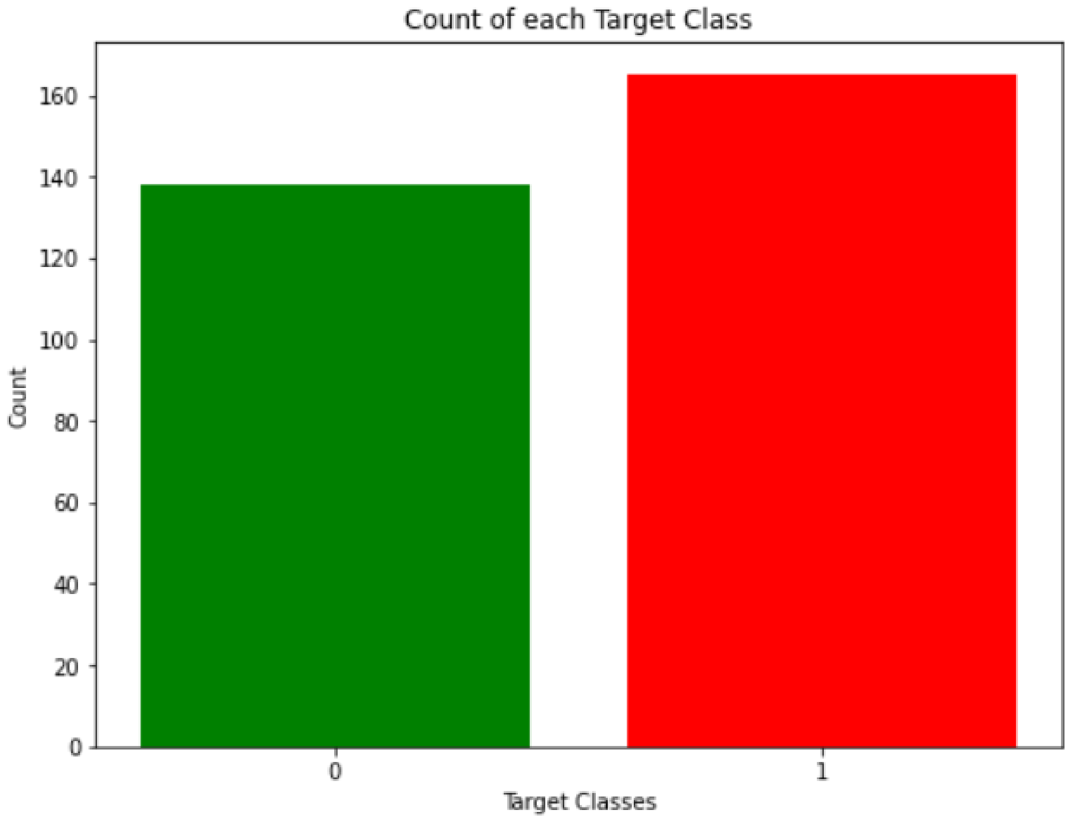


Outcome (Heart Attack & No Heart Attack) against Exercise Induced Angima

5. Data Split

In this project, data is split based on a ratio of 80:20 for the training set and the test set. The training set data is used in the logistic regression component for model training, while the prediction set data is used in the prediction component.

The Dataset class balance is as below:



D) Solution Statements

- To make a Logistic Regression Model (Sagemaker Linear Learner), the problem being a binary classification with very less correlation between features.
- To deploy the trained model on AWS Sagemaker and subsequently deploying an endpoint which can be used to make predictions

E) Benchmark Model

For benchmarking, I will pick a Random Forest model against my solution. The predicted labels will be compared to the original labels to find false positives and false negatives. Number of false positives and false negatives will tell us about the performance of the model.

F) Evaluation Metrics

The model will be using various evaluation metrics such as

- Accuracy: which refers to how close a measurement is to the true value and can be calculated using the following formula

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

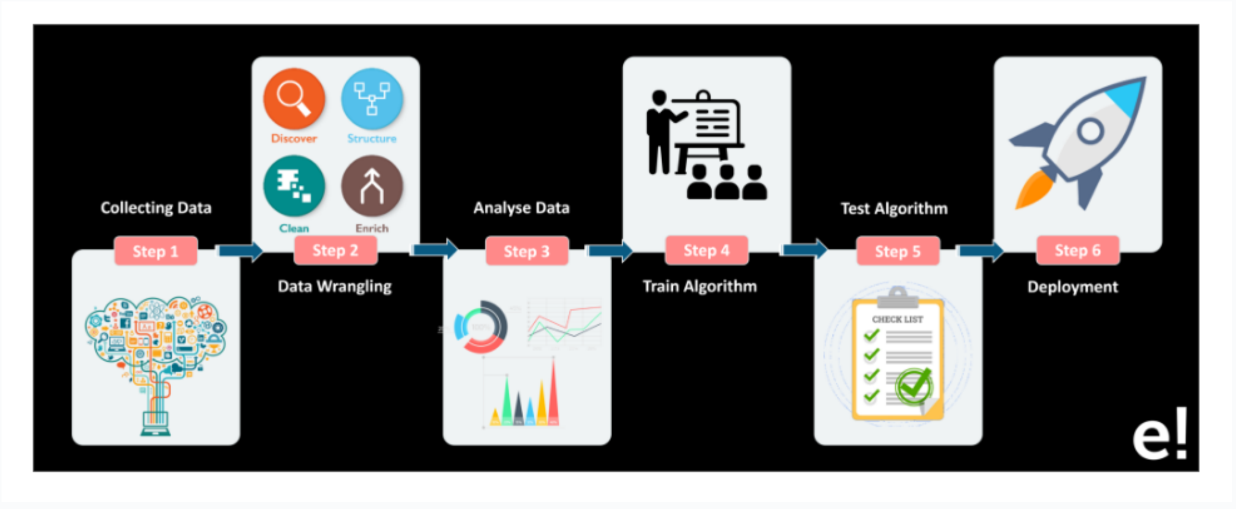
- Precision: which is how consistent results are when measurements are repeated and can be calculated using the following formula

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

- Recall: which refers to the percentage of total relevant results correctly classified by the model and can be calculated using the formula

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

G) Project Design



To develop a ML model successfully, we must go through the following stages:

The diagram above gives an overview of how our ML model will work. On the far right is Step-1 - Data Collection, in this case I will be using UCI Heart Disease dataset from kaggle.

Step-2 and Step-3 - Data pre-processing which includes feature scaling and analyzing data. This will be done using Panda and various Python in-built libraries.

Step-4 - Training Algorithm. When the dataset is split into two parts - Training set & Test set, an appropriate Machine Learning algorithm (Linear-Learner) is applied. Training set is the one through which the model learns. Test set is used for calculation the accuracy, precision, recall etc. and the Training data is fed into the algorithm after defining the parameters of ML algorithm.

Step-5 Test Algorithm. After the model is trained, it will be tested with Test dataset. If the results are not satisfactory, the model should be reconfigured by varying the parameters (Hyperparameter Tuning).

Step-6 Deployment. The ML trained model will be deployed using AWS Sagemaker and deploying an endpoint which can be used to make predictions.

.