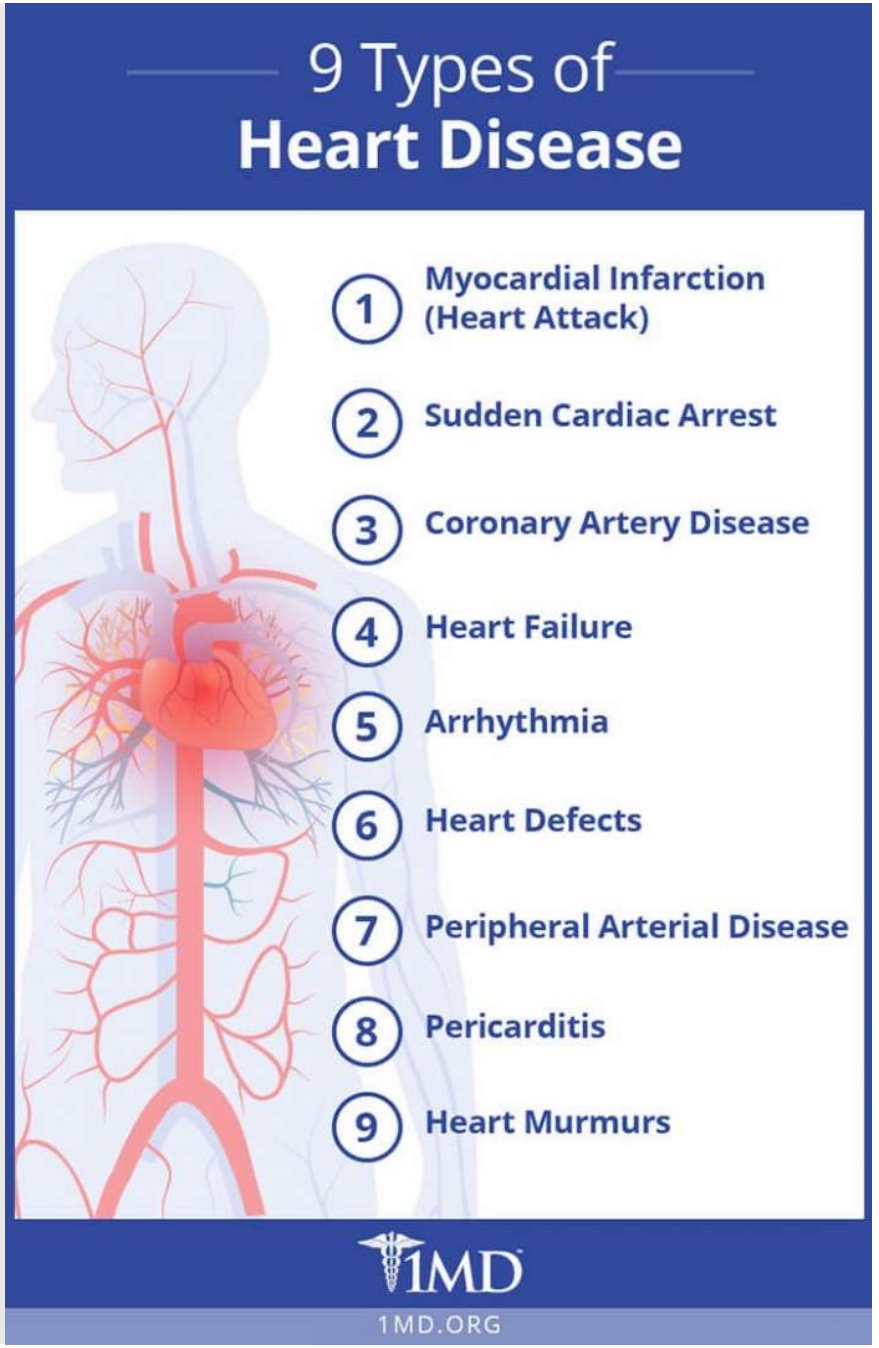
PREDICTING HEART DISEASE WITH MACHINE LEARNING

By Tony DiRubbo

PROPOSAL & FEATURE ENGINEERING

What is Heart Disease?

- Many different names
- Disease which directly impacts the heart or blood vessels
- The leading cause of death in the United States
- Research questions:
  - Classification – Can a Heart Disease diagnoses be predicted by various numerical variables
  - Regression – Can we predict the number of colored vessels in a fluoroscopy



Feature Engineering

- UC Irvine Machine Learning Repository
  - 920 Observations
    - Observations with NA values will have the median inputted
  - 14 Variables
    - All variables will be split into different groups for appropriate analyses

Install the ucimlrepo package

```
pip install ucimlrepo
```

Import the dataset into your code

```
from ucimlrepo import fetch_ucirepo

# fetch dataset
heart_disease = fetch_ucirepo(id=45)

# data (as pandas dataframes)
X = heart_disease.data.features
y = heart_disease.data.targets

# metadata
print(heart_disease.metadata)

# variable information
print(heart_disease.variables)
```

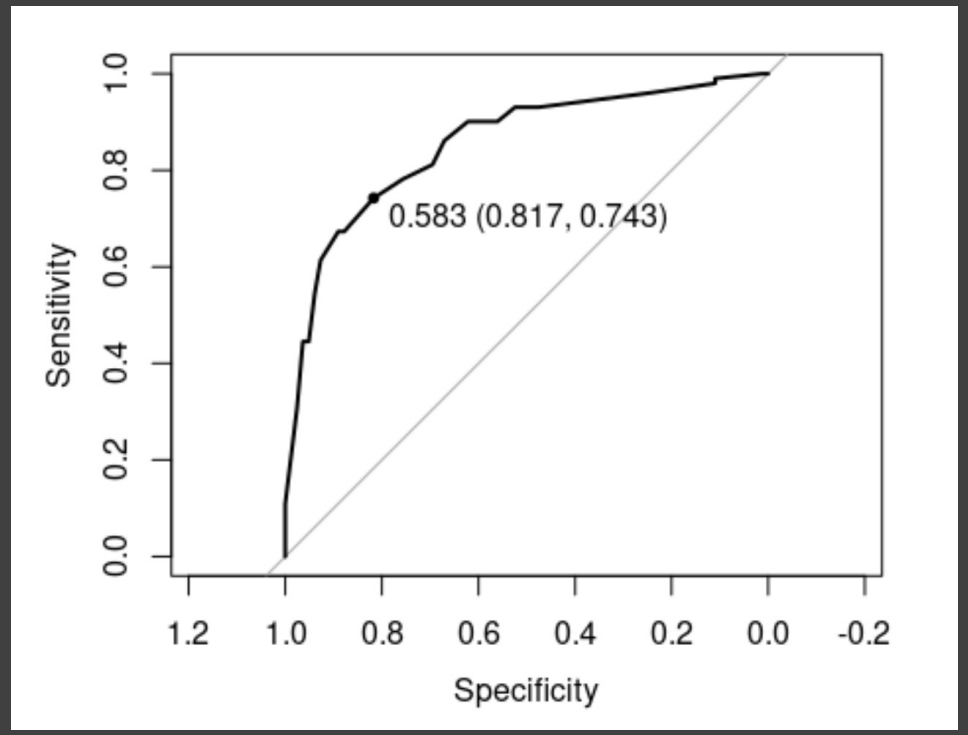
```
HDdf <- read_csv("heart_disease_uci.csv") #reading in the CSV

#Updates data so that heart disease diagnoses is a two factor variable
HDdf <- within(HDdf,{num<-factor(num, labels = c("No","Yes","Yes","Yes","Yes"))})

#renaming variables so that they are easier to understand at a quick glance
names(HDdf)[names(HDdf) == "cp"] <- "CerebralPalsy"
names(HDdf)[names(HDdf) == "chol"] <- "Cholestoral"
names(HDdf)[names(HDdf) == "fbs"] <- "FastingBloodSugar"
names(HDdf)[names(HDdf) == "restecg"] <- "RestingECGReading"
names(HDdf)[names(HDdf) == "exang"] <- "ExercisedAngina"
names(HDdf)[names(HDdf) == "trestbps"] <- "RestingBPS"
names(HDdf)[names(HDdf) == "thalch"] <- "MaxHeartRate"
names(HDdf)[names(HDdf) == "oldpeak"] <- "STDepression"
names(HDdf)[names(HDdf) == "slope"] <- "STSlope"
names(HDdf)[names(HDdf) == "ca"] <- "FluoroscopyColoredVessels"
names(HDdf)[names(HDdf) == "thal"] <- "ThalassemiaDiagnoses"
names(HDdf)[names(HDdf) == "num"] <- "HeartDisease"

#saves data as a data frame
HDdf <- as.data.frame(HDdf)
```

```
HDdf <- HDdf %>%
  mutate(age = ifelse(is.na(age), median(age, na.rm=TRUE), age),
         RestingBPS = ifelse(is.na(RestingBPS), median(RestingBPS, na.rm=TRUE), Res
         Cholestoral = ifelse(is.na(Cholestoral), median(Cholestoral, na.rm=TRUE),
         MaxHeartRate = ifelse(is.na(MaxHeartRate), median(MaxHeartRate, na.rm=TRUE
         STDepression = ifelse(is.na(STDepression), median(STDepression, na.rm=TRUE
         FluoroscopyColoredVessels = ifelse(is.na(FluoroscopyColoredVessels), media
```

```
#NumericVariablesOnly
HDn <- select(HDdf, c("age","RestingBPS","Cholestoral","MaxHeartRate","STDepressior

#RegressionVariables
HDreg <- select(HDdf, c("HeartDisease","age","RestingBPS","Cholestoral","MaxHeartRa

#Creating Dummy Variables for Regression (HDreg2)
#ifelse(test_expression, x, y)
HDreg$HeartDisease<-ifelse(HDreg$HeartDisease == "Yes",1.0,0)
names(HDreg)[names(HDreg) == "sex"] <- "Male"
HDreg$Male<-ifelse(HDreg$Male == "Male",1,0)
HDreg$FastingBloodSugar<-ifelse(HDreg$FastingBloodSugar == TRUE,1,0)
names(HDreg)[names(HDreg) == "RestingECGReading"] <- "Hypertrophy"
HDreg$Hypertrophy<-ifelse(HDreg$Hypertrophy == "normal",0,1)
HDreg$ExercisedAngina<-ifelse(HDreg$ExercisedAngina == TRUE,1,0)
HDreg$ThalassemiaDiagnoses<-ifelse(HDreg$ThalassemiaDiagnoses == "normal",0,1)

# Impute missing values with mean column-wise
for (col in names(HDreg)) {
  HDreg[is.na(HDreg[, col]), col] <- mean(HDreg[, col], na.rm = TRUE)
}
```

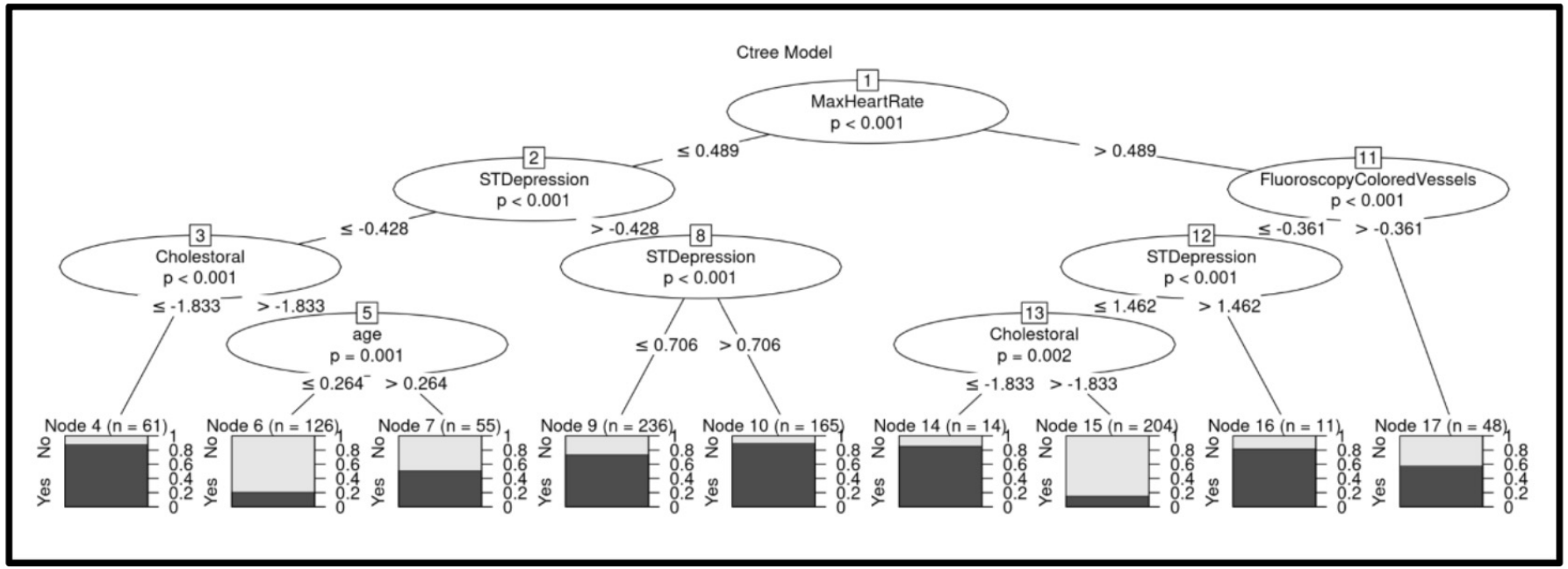
PROGRESS REPORT 1

Creating Baseline Models

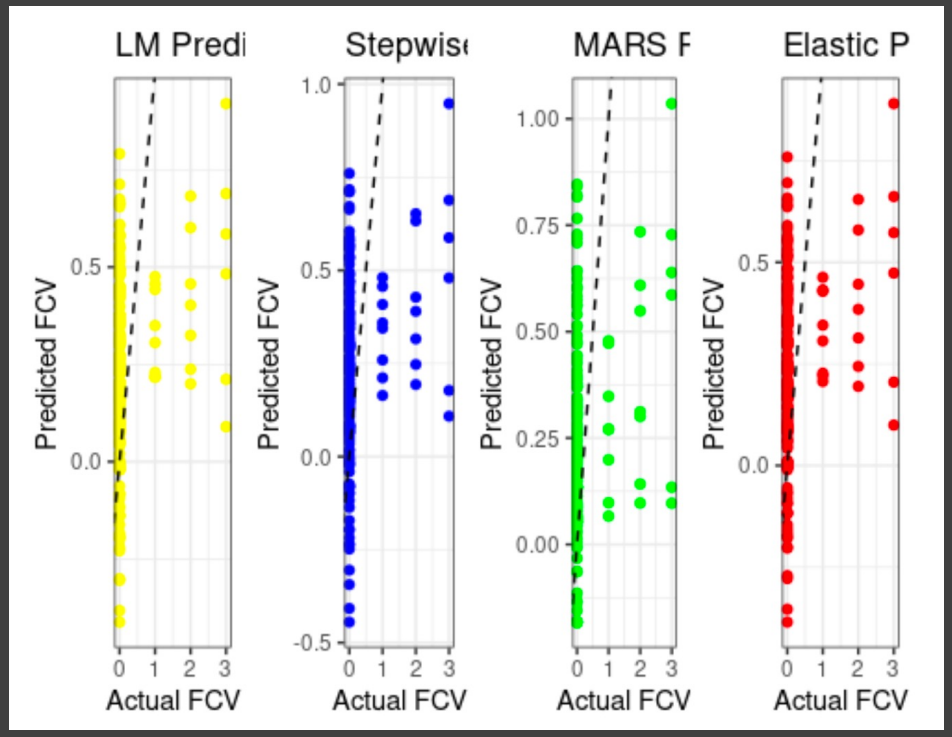


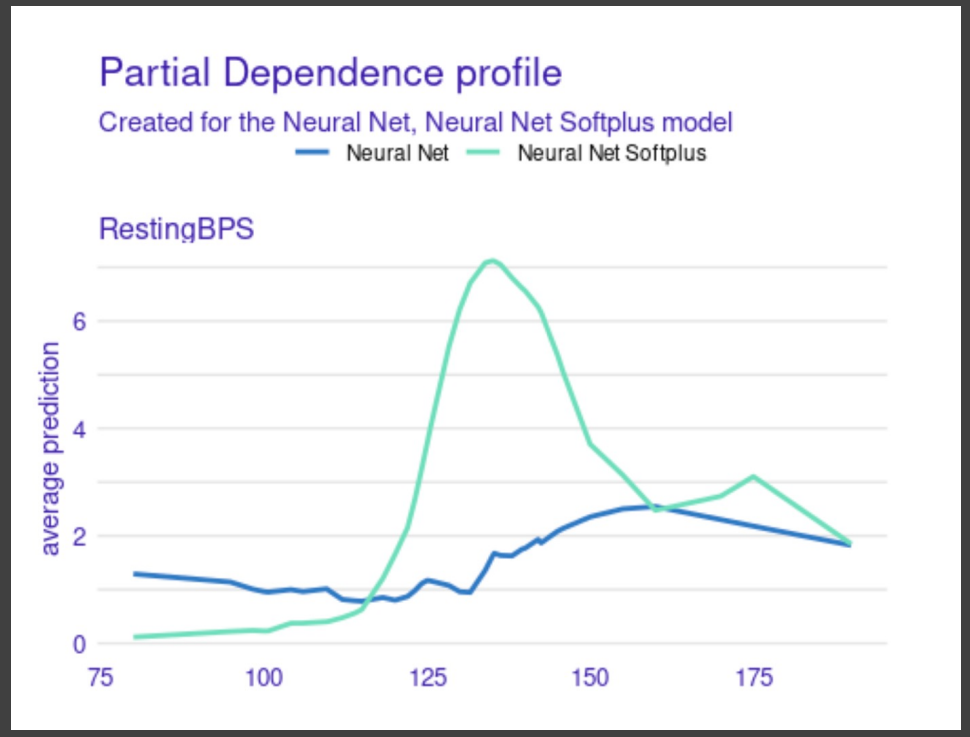
	k	Accuracy	Kappa	AccuracySD	KappaSD
1	1	0.7014837	0.3930920	0.01732227	0.03726104
2	3	0.7462352	0.4831252	0.01878982	0.03768980
3	6	0.7529644	0.4984139	0.02141121	0.04389764
4	9	0.7597488	0.5114171	0.02389708	0.04972416
5	12	0.7570274	0.5065319	0.02352874	0.04800120
6	15	0.7705503	0.5334751	0.03357268	0.07031367
7	18	0.7814533	0.5570762	0.02335190	0.04869017
8	21	0.7556393	0.5051209	0.02934668	0.06150452
9	24	0.7705410	0.5349989	0.03813845	0.07986705
10	27	0.7597118	0.5127321	0.03176209	0.06563591
11	30	0.7651449	0.5241996	0.02742231	0.05794253

MODEL 1: KNOWN NEAREST NEIGHBOR



MODEL 2: CLASSIFICATION TREE





MODEL 3: FOUR REGRESSIONS COMPARED
MODEL 4: ARTIFICIAL NEURAL NETWORK

PROGRESS REPORT 2

Addressing Regression

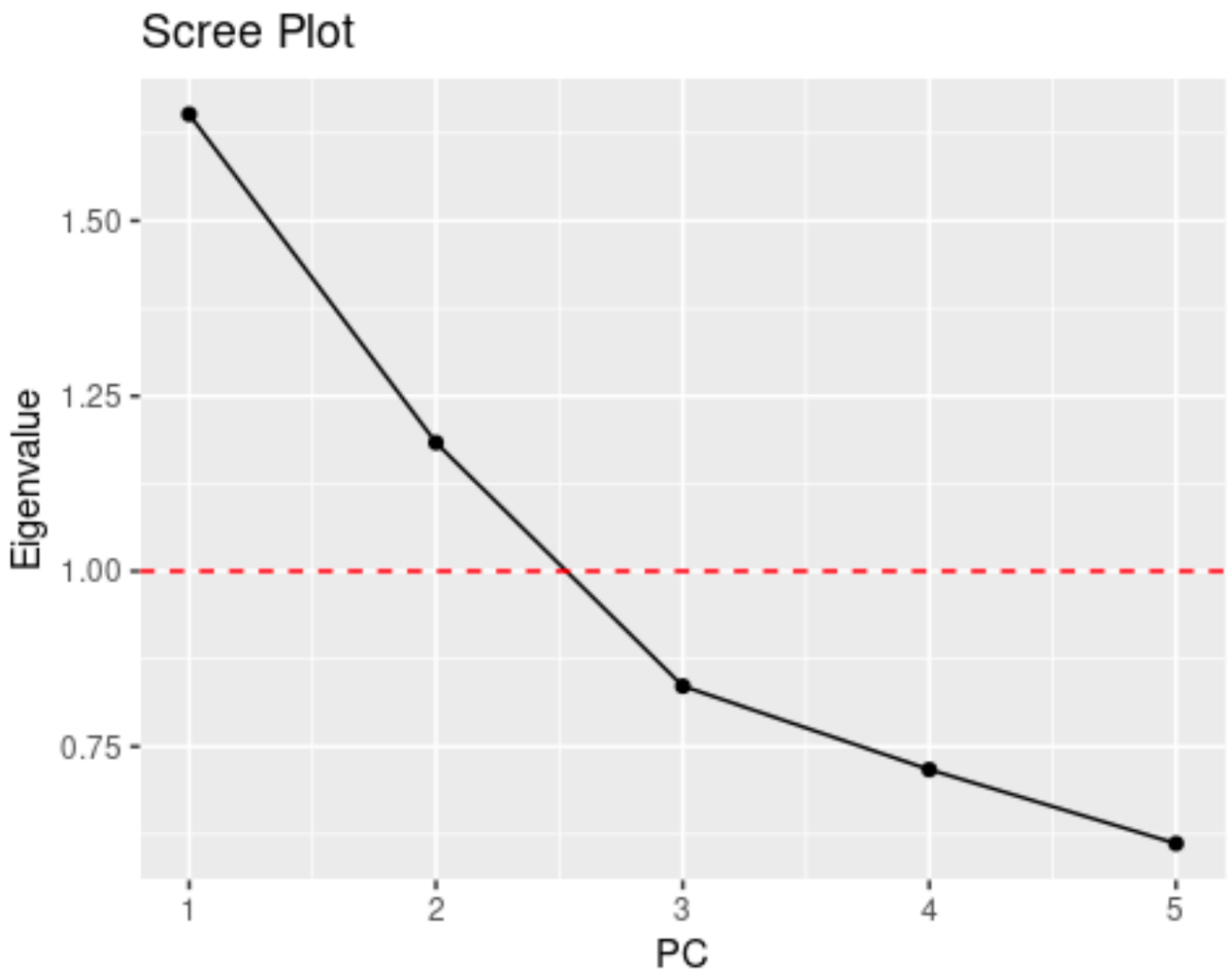
MAIN PROBLEM:
REGRESSION PERFORMANCE LAGGING CLASSIFICATION PERFORMANCE

Performance of Regression ML Models

	Performance Measure (Value for Best Fit)		
Model Type	**MAE (0)**	**RMSE(0)**	**R-Squared (1)**
Basic Linear Regression	0.369	0.572	0.130
Stepwise Regression	0.371	0.581	0.123
MARS Regression	0.364	0.588	0.106
Elastic Regression	0.369	0.580	0.133
Logistic Activation ANN	0.315	0.626	0.012
SoftPlus Activation ANN	0.276	0.697	0.163

Solution 1: PCA

- Principle Component Analysis – Creating new variables (components) from previous variables to separate aspects of correlation
  - Used primarily with Likert scale responses
  - Scree plot only wanted two components to be used, which would not have been enough.



SOLUTION 2: DUMMY VARIABLES

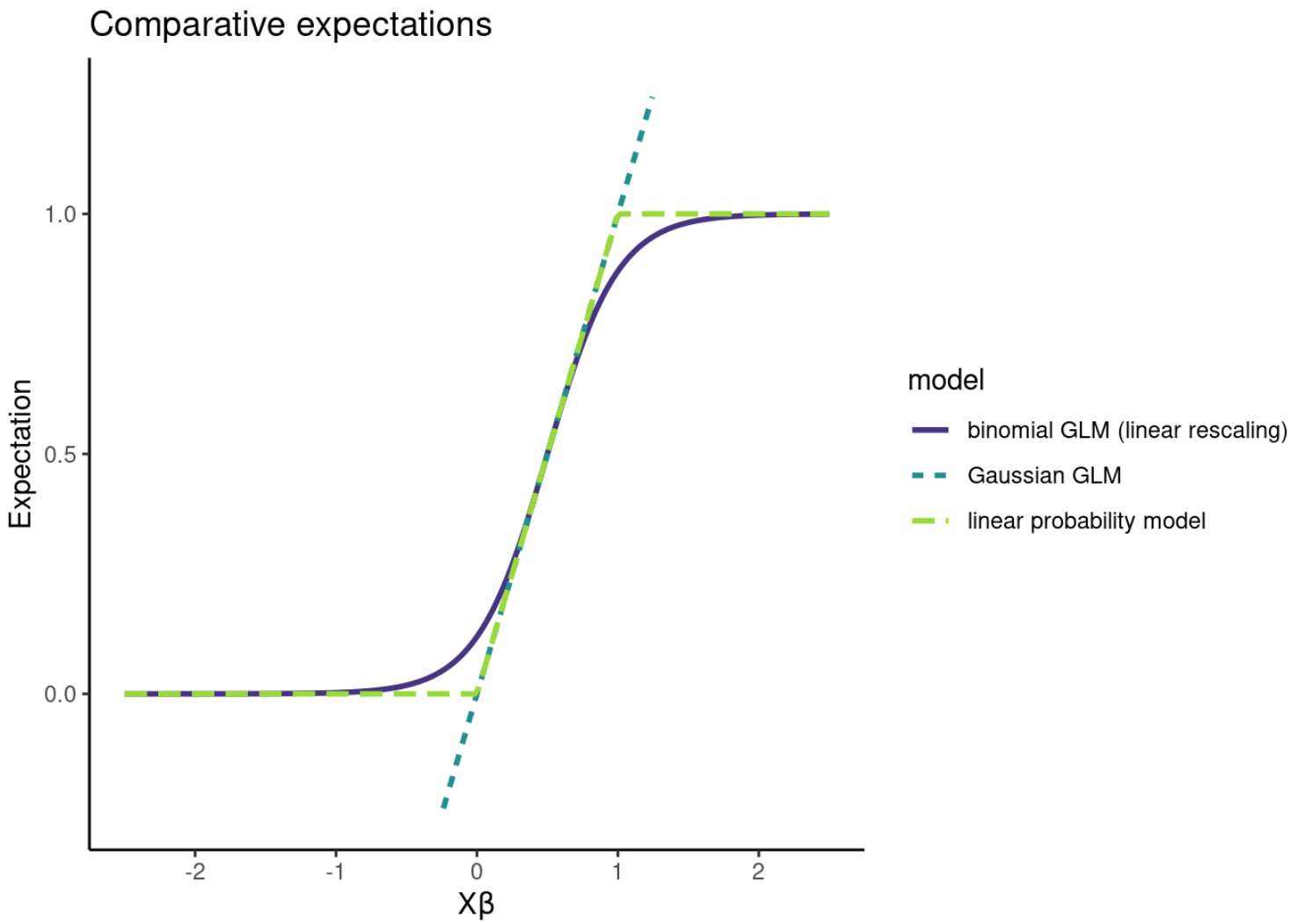
Updated Performance of Regression ML Models

New Regression -> Regression with Dummy Variables

	Mean Performance Measure (Value for Best Fit)		
Model Type	**MAE (0)**	**RMSE(0)**	**R-Squared (1)**
Basic Linear Regression	0.369	0.572	0.130
New Basic Linear Regression	0.370	0.599	0.089
Stepwise Regression	0.371	0.581	0.123
New Stepwise Regression	0.374	0.600	0.073
MARS Regression	0.364	0.588	0.106
New MARS Regression	0.355	0.570	0.163
Elastic Regression	0.369	0.580	0.133
New Elastic Regression	0.364	0.596	0.080
Logistic Activation ANN	0.315	0.626	0.012
New Logistic Activation ANN	0.447	0.905	0.002
SoftPlus Activation ANN	0.276	0.697	0.163
New Softplus Activation ANN	Dummy Variables Cause Model to Crash		

Linear Probability Model (LPM)

- Turn "Heart Disease" into a dummy variable
- Make this variable the new dependent variable
- Regression now calculates a predicted value between 0 and 1 (0% and 100%)
- Machine learning regression models are now predicting percent chance that individual develops heart disease
  - Classification model predicts if you have heart disease
  - Regression model predicts the chance you develop heart disease
- Caret Ensemble and caretList() allow for many regressions to be quickly compared



FINAL PORTFOLIO

adirubbo13.github.io