CIND 820

Determinants of Financial Well-Being

Capstone Project Course
Ryerson University

Initial Results

By

Zeynep B Cevik
501138112

Supervised By

Uzair Ahmad, Ph.D.

November 8, 2021

4. Initial Results

In this section, the initial results of the models that are constructed for this study will be explained.[1]

The research questions of this study are as follows:

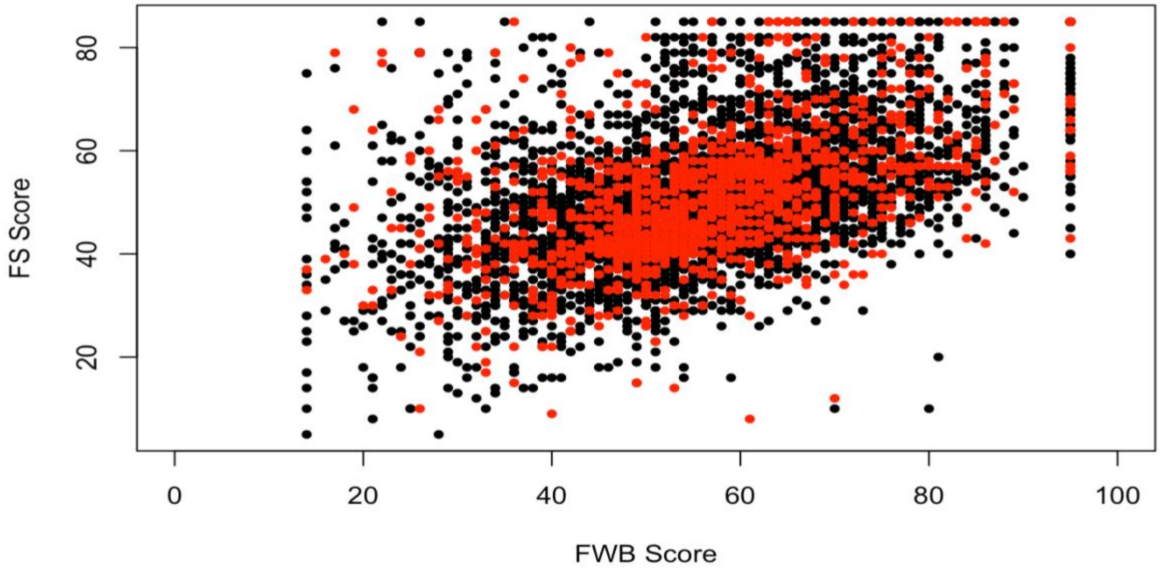
- Which factor(s) influences an individual's well-being?
- If the demographic structure or various socioeconomic factors are controlled, will the financial behaviors such as financial skills be still significant in determining the financial well-being?
- Is there any factor that may work together to determine an individual's level of financial well-being or which factors are more effective?

To answer these questions, a score for FWB, which is calculated by CFPB, is used as explained in the previous sections in detail. Various machine learning methods are used for the search of above-mentioned questions. These methods are Decision Tree Method, Logistic Regression, Ordinary Least Square Regression, k-NN classification and regression and lastly Naïve Bayes Method.

The score of FWB is a discrete variable between 0 and 100. To be able to use this score in logistic regression or k-NN classification and Naïve Bayes Method, this score is returned into two different categorical variables[2], one is binary and the other is between 1 and 4, such as:

$$\text{FWB\_cat} = \begin{Bmatrix} 1\, if\ 0 < FWB \leq 25 \\ 2\, if\ 25 < FWB \leq 50 \\ 3\, if\ 50 < FWB \leq 75 \\ 4\, if\ 75 < FWB \leq 100 \end{Bmatrix} \quad \text{and FWBcat\_bin} = \begin{Bmatrix} 1\, if\ 0 < FWB \leq 50 \\ 2\, if\ 50 < FWB \leq 100 \end{Bmatrix}$$

Figure 1: Training and Test Datasets (Black Dots: Training Dataset, Red Dots: Test Dataset)



[1] R-programming language is used throughout initial results section. The detailed version of how I found the solutions are in R-Markdown file.

[2] The results for this categorical variable are not presented in this study, but can be found in the codes file, because the values 1 and 4 are never used in the methods. So, the binary categorical variable is demonstrated in this study.

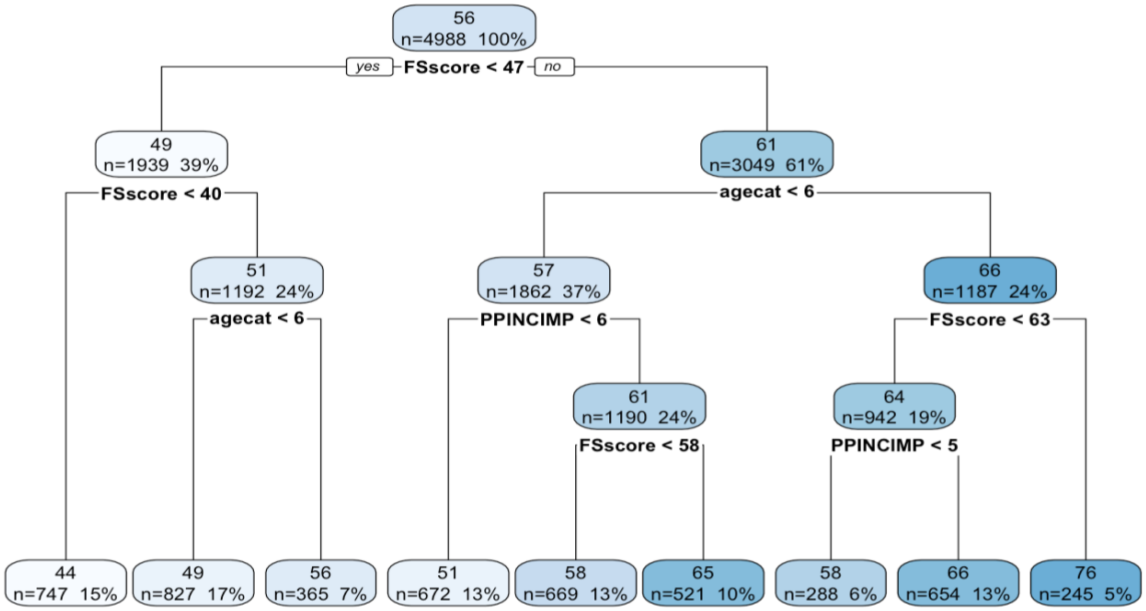
To be able to evaluate the model performance such as accuracy, sensitivity, etc., the dataset is divided into two: training and test datasets. (0.80/0.20 ratio is used.) As an example of the distribution of these two datasets, a scatter plot between FWB score and Financial Skill (FS) score is shown in Figure 1.

4.1. Decision Tree Method:

The decision tree method is a widely used predictive analytic tool among ML algorithms. The tree is constructed in a recursive partitioning way so that each internal node has a threshold level, so that at the end all the observations are included. The decision tree method is used to help to identify the significant variables in explaining the dependent variable.

Figure 2 represents the decision tree for FWB score, which is between 0 and 100. According to this figure, FS score, age[3], and income are significant determinants of FWB. The first threshold level in FS score is 47, meaning that if an individual has a lower than 47 points FS score, there is a very high probability that individual has a low FWB score. The individuals who have the highest FWB scores, have a FS score more than 63 and bigger than 69 years of age. The lowest FWB scores belong to the individuals who have lower than 40 FS score. FS can be interpreted as financial knowledge. So, this figure shows that if an individual has a low financial knowledge, there is a very high probability that he/she has a low FWB. For instance, you have a high financial knowledge, and your age is less than 45 (category<3), if your income is less than $60,000, then the probability of having a high FWB is very low. On the other hand, if your income is higher than 60,000 then you may have a higher-than-average FWB score.

Figure 2: Decision Tree Outcome for Discrete Dependent Variable (FWB)



[3] As explained in the data description section, age and income are categorical variables. Category 6 for age is between 62-69, category 5 for income is annual income between 50,000 and 60,000 and category 6 is between $60,000 and $75,000.

Figure 3: The Residuals = Actual Value-Predicted Value

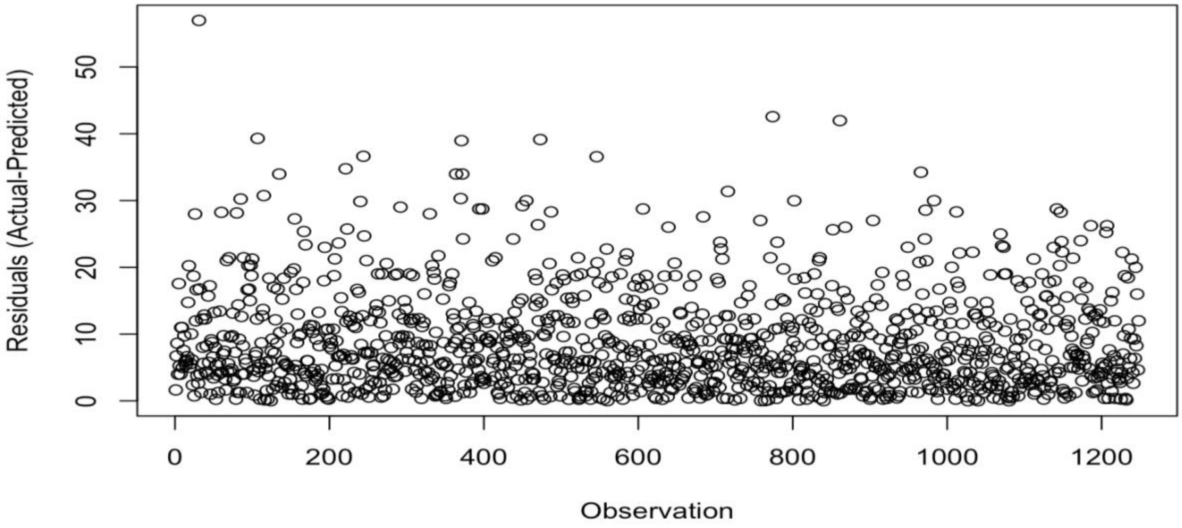
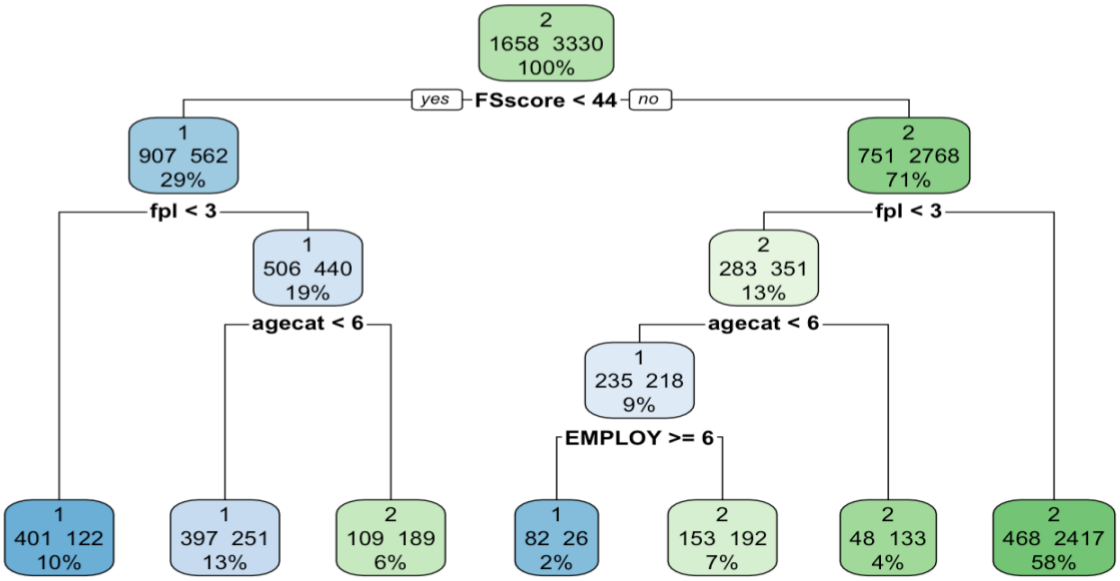


Figure 4: Decision Tree Outcome for Categorical Binary Dependent Variable (FWBcat_bin)



In Figure 3, the residuals, which are calculated by the difference between actual and fitted values, are shown. The sign of having a good model is that the residuals should be close to the value zero. In this case, I can say most of the residual values are close to zero, however there are also many positive values which imply that I need to look at other model metrics.

When FWB is divided into two categories[4]: low (1) and high (2), the outcome slightly changes. FS score and age are still two significant independent variables, in addition to those, poverty and employment status are also important in determining FWB. Figure 4 shows the related decision tree outcome. If your financial knowledge is more than the average and you earn at least twice the poverty level[5], then your FWB score is more than 50. So, poverty status can also be interpreted as

[4] It is 1 if FWB is less than or equal to 50, 2 if it is higher than 50.

[5] The poverty level in 2021 for United States is $26,500 for a household of 4 individuals.

income status of that individual. The employment and age variables play role in determining FWB, when the individual earns less than twice poverty level.

Table 1. Confusion Matrix and Model Specific Metrics for Categorical Binary Dependent Variable

Confusion Matrix	**Actual**	
Prediction	**1**	**2**
1	880	399
2	778	2931

Accuracy	0.764
95%CI	(0.752, 07758)
No Information Rate	0.6676
P Value	<2.2.e-16
Kappa	0.436
Mcnemar's Test P-Value	<2.2 e-16
Sensitivity	0.5308
Specificity	0.8802
Pos Prediction Value	0.6880
Neg Prediction Value	0.7902
Prevalence	0.3324
Detection Rate	0.1764
Detection Prevalence	0.2564
Balanced Accuracy	0.7955

The confusion matrix gives information about the model performance and model metrics such as accuracy, sensitivity, etc. are calculated by using confusion matrix. For instance, accuracy shows the number of correct predictions over all predictions. In this model, 0.76 accuracy means that with 76% of the predictions are accurate. However, I also need to check the other evaluation measures too. When I look at the sensitivity and specificity, I can conclude that the correctly predicted positive data points is way too less than the correctly predicted negative data points. Sensitivity is equal to 0.53, meaning 53% of positive data points are correctly predicted. On the other hand, specificity is 0.88, meaning 88% of negative data points are correctly predicted. So, if I am interested in seeing the prediction performance for negative data points, this model is a good model. Furthermore, the very low p-value demonstrates the significance of the model. In other words, the model is statistically significant and can be used for predictions.

4.2. Logistic Regression Method:

Logistic regression is a kind of regression when the dependent variable is a categorical variable. In table 2, there are two models. The first model includes 21 variables whereas in the second model, only the statistically significant variables are included. The results are pretty much the same. There are only slight differences both in the model performance and in the coefficients. Although the first model gives a little higher Akaike-Schwartz Criteria (AIC), it also demonstrates slightly higher R-squared and lower residual level. The coefficients do not also change drastically. Hence, I will only interpret the coefficients of the first model.

According to the first model results, financial knowledge, age, ethnicity, income, and marital status are statistically significant in determining the FWB. The most significant attribute is financial knowledge and as financial knowledge increases, FWB also increases. Also, as the individual gets older and older and earns more and more, FWB increases. On the other hand, as the ethnicity goes from white, non-Hispanic to non-white, FWB decreases.

The goodness of fit of the model is low (0.24 approximately). It means that 24% of the FWB variations can be explained by the variations of this model's independent variables. So, we need to add more relevant variables to increase the explanation power of the model.

Table 2. Results of Logistic Regression Models[6]

Dependent Variable: **FWB Binary Categorical Var.**	**Model (1)**	**Model (2)**
Intercept	-4.5269*** (< 2e-16)	-5.1124*** (< 2e-16)
FS Score	0.0718*** (< 2e-16)	0.0739*** (< 2e-16)
KH Score	0.4030*** (1.64e-10)	0.3843*** (9.07e-14)
Age	0.2186*** (< 2e-16)	0.2566*** (< 2e-16)
Ethnicity	0.0957** (0.0069)	0.0881** (0.0095)
Income	0.2441*** (< 2e-16)	02191** (< 2e-16)
Marital Status	-0.0973*** (0.0008)	-0.091** (0.0095)
Poverty	-0.0718 (0.3967)	
LM Score	-0.0482 (0.4032)	
Employment	0.0211 (0.2304)	
Military Status	-0.0086 (0.8259)	
Education	0.0112 (0.7633)	
Gender	-0.0078 (0.9180)	
Household size	-0.1055 (0.3399)	
MSA Status	-0.0201 (0.8473)	
Census Region	0.0476 (0.1798)	
Presence of Household Members – Children 0-1	0.0462	

[6] P-values are in parentheses

	(0.8297)	
Presence of Household Members – Children 2-5	0.1014 (0.5641)	
Presence of Household Members – Children 6-12	0.0256 (0.8821)	
Presence of Household Members – Children 13-17	-0.1219 (0.4465)	
Presence of Household Members – Adults 18+	-0.0196 (0.8671)	
R-Squared	0.2469	0.2440
AIC	4819.1	4809.8
Residual Deviance	4777.1	4795.8

***Significant at 99.9% Confidence Level, ** Significant at 99% Confidence Level

Table 3. Confusion Matrix and Model Specific Metrics for the First Model

Confusion Matrix	**Actual**	
Prediction	**1**	**2**
1	223	116
2	195	714

Accuracy	0.7508
95%CI	(0.7528, 0.7746)
No Information Rate	0.6651
P Value	2.95.e-11
Kappa	0.4131
Mcnemar’s Test P-Value	9.735e-06
Sensitivity	0.5335
Specificity	0.8602
Pos Prediction Value	0.6578
Neg Prediction Value	0.7855
Prevalence	0.3349
Detection Rate	0.1787
Detection Prevalence	0.2716
Balanced Accuracy	0.6969

4.3 Ordinary Least Squares Regression (OLS) Method:

Ordinary Least Squares Regression is a method of regression, aiming to minimize the square of residuals. For OLS method, the discrete value of FWB is used. Table 4 shows the outcome of OLS method. There is one model because forward selection and backward elimination algorithms give the most appropriate model with the highest AIC values.

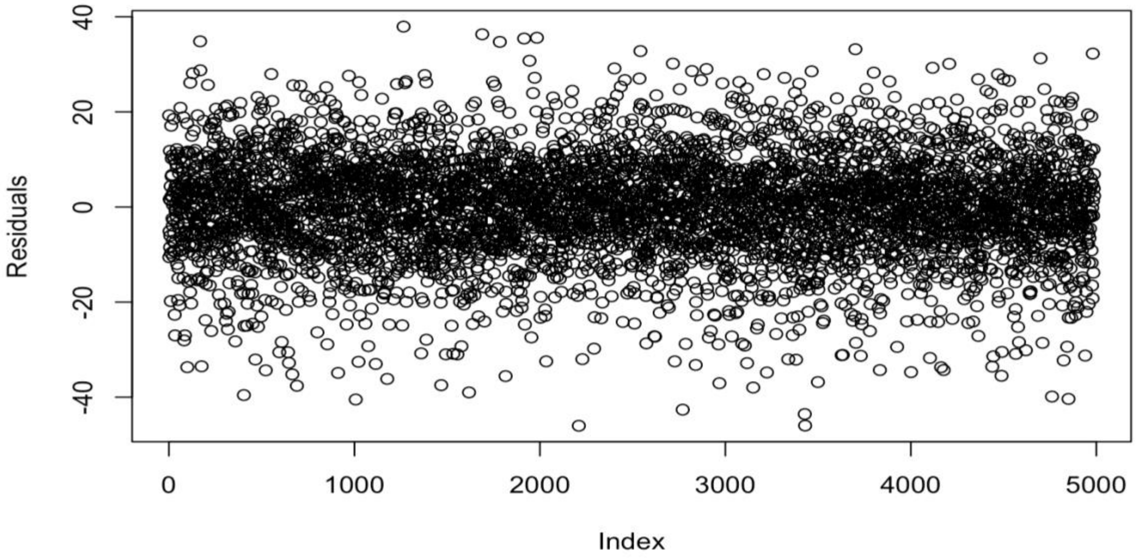
According to the OLS model results, with 90% confidence, financial knowledge, age, ethnicity, income, household size, census region, presence of 13-17 years old children, marital and employment statuses are statistically significant. To be more specific, as the financial skill score increases by 1 point, FWB score increase by 0.45 point. As age and income increase, FWB also increases. However, as the household size increases, FWB decreases. The same result with the ethnicity is the same with the logistic regression: as the ethnicity goes from white, non-Hispanic to non-white, FWB decreases. One of different outcomes from the logistic regression is the significant effect of a teenager in the household. If the household has a teenager in the household, FWB is negatively affected. The adjusted R-squared show the goodness of the fit. In this model it is approximately 0.45 which means that this model with these independent variables can only explain approximately the half of the variations seen in FWB score. The low level of p-value confirms that the model is statistically significant since it means rejecting the null hypothesis which is the model outcome is not statistically significant. In addition, as shown in Figure 5, the residuals of this model are heavily on the zero line which implies most of the time the fitted value and the actual values are so close to each other.

Table 4. Results of OLS Model

Dependent Variable: **FWB Score**	**Model**	**P-Values**
Intercept	19.5441***	< 2e-16
FS Score	0.45278***	< 2e-16
KH Score	1.3558***	7.44e-12
Age	1.2962***	< 2e-16
Ethnicity	0.2790*	0.0409
Income	1.4789***	< 2e-16
Marital Status	-0.2947*	0.0103
Employment	0.3682***	1.173e-08
Military Status	-0.1962	0.1487
Household size	-0.6192***	1.31e-05
Census Region	0.3062*	0.0222
Presence of Household Members – Children 13-17	-0.8509 `	0.0762
Adj. R-Squared	0.4237	
p-Value	< 2.2e-16	
Residual Standard Error	10.69	

***Significant at 99.9% Confidence Level, ** Significant at 99% Confidence Level, *95% Confidence Level, `90% Confidence Level

Figure 5. Residuals of OLS Model



4.4. k-NN Classification and Regression:

In this section, k-NN regression and classification algorithms will be applied. Key difference between k-NN classification and regression is that "*KNN regression tries to predict the value of the output variable by using a local average. KNN classification attempts to predict the class to which the output variable belongs by computing the local probability*"[7] So, in k-NN classification, the dependent variable is a categorical variable, whereas in k-NN regression model, it is a quantitative variable, in this study FWB score is discrete variable.

Figure 6. Optimal K Value in k-NN Regression -Elbow Method

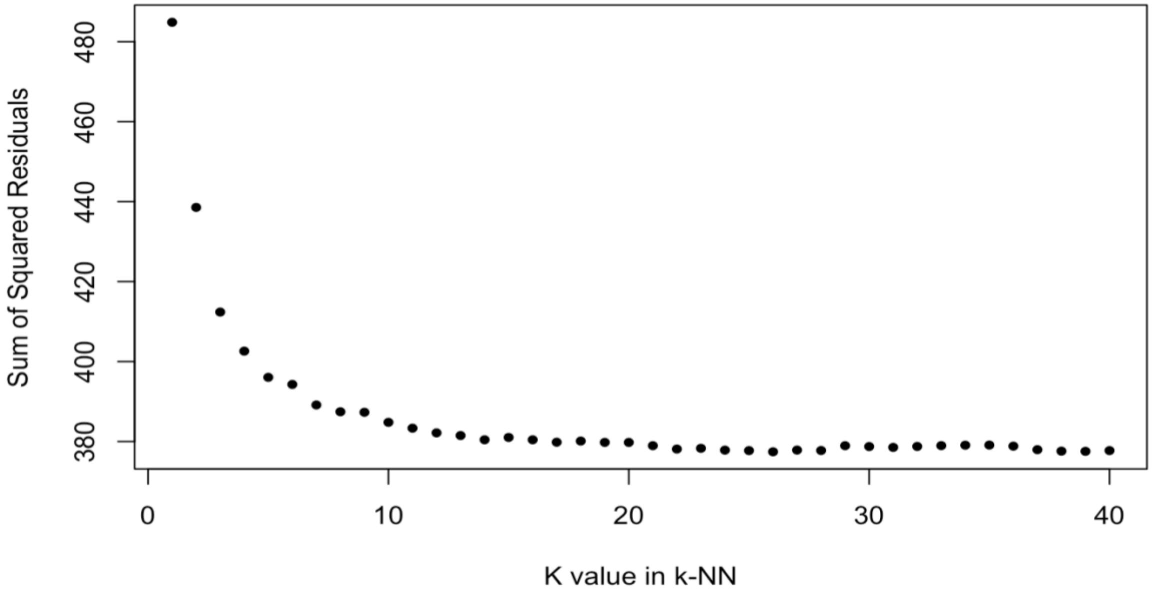


Table 5. Confusion Matrix and Model Specific Metrics for the k-NN Classification (k=15)

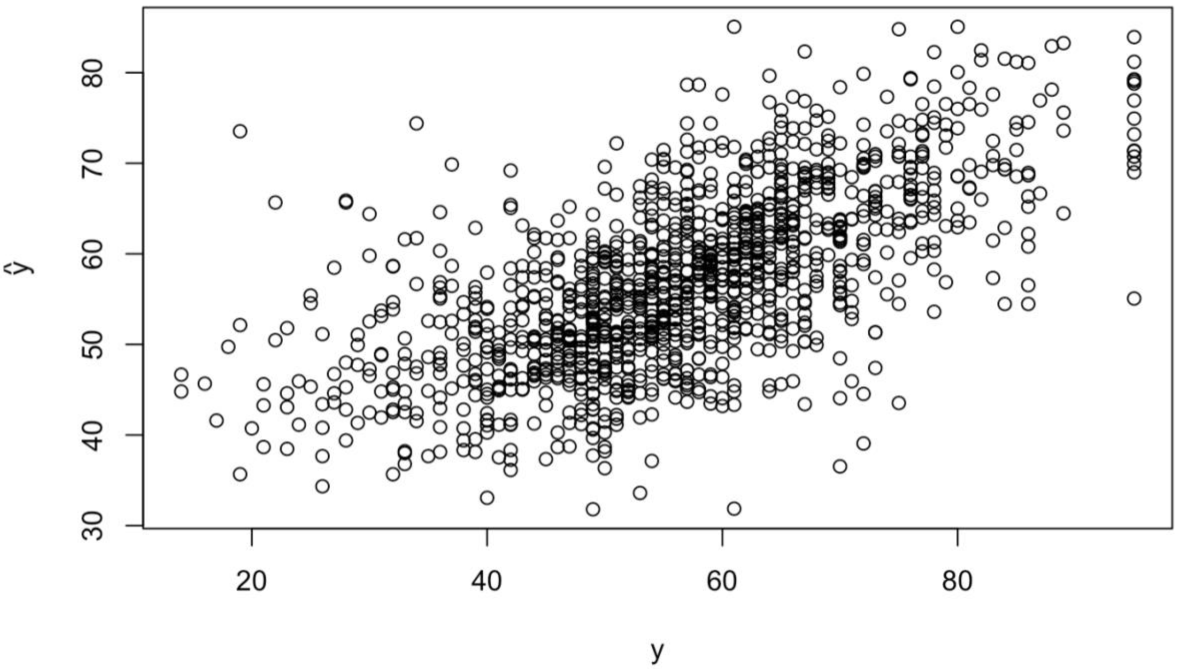
Confusion Matrix	**Actual**	
Prediction	**1**	**2**
1	224	119
2	194	711

Accuracy	0.7492
95%CI	(0.7242, 0.773)
No Information Rate	0.6651
P Value	6.832.e-12
Kappa	0.4108
Mcnemar's Test P-Value	2.88e-06
Sensitivity	0.5359
Specificity	0.8566
Pos Prediction Value	0.6531
Neg Prediction Value	0.7856
Prevalence	0.3349

[7] This phrase is cited from https://stats.stackexchange.com/questions/364351/regression-knn-model-vs-classification-knn-model

Detection Rate	0.1795
Detection Prevalence	0.2748
Balanced Accuracy	0.6963

Figure 7. Fitted Values Versus Actual Values for k-NN Regression (k=15)



One of the important steps is deciding the number of clusters (k). In the literature, there is no consensus about finding the number of clusters- k. The suggested method is known as Elbow Method which is based on error rate. The less the error rate, the better the model is. Figure 6 shows the sum of squared residuals for each number of clusters starting from 1 to 40. There is a steady decrease in the sum of squares as k increases. However, the decrease becomes less and less after k=15. So, k=15 can be said to be optimal in this case. I use k=15 for this section.

In Table 5, the confusion matrix and specific model metrics for k-NN Classification method is presented. In k-NN classification method, the FWB, which is converted into binary categorical variable, is used. The confusion matrix gives information about the model performance and model metrics such as accuracy, sensitivity, etc. are calculated by using confusion matrix. For instance, accuracy shows the number of correct predictions over all predictions. In this model, 0.75 accuracy means that with 75% of the predictions are accurate. However, I also need to check the other evaluation measures too. When I look at the sensitivity and specificity, I can conclude that the correctly predicted positive data points is way too less than the correctly predicted negative data points. Sensitivity is equal to 0.53, meaning 53% of positive data points are correctly predicted. On the other hand, specificity is 0.86, meaning 86% of negative data points are correctly predicted. So, if I am interested in seeing the prediction performance for negative data points, this model is a good model. Furthermore, the very low p-value demonstrates the significance of the model. In other words, the model is statistically significant and can be used for predictions.

Figure 7 demonstrates a scatter plot between the fitted value, which resulted from k-NN regression, versus the actual data. So, if most of the dots lie on the diagonal line (x=y line), then it signals that the error is small, and the model can give accurate predictions. In this example, I can say most of the dots lie on the diagonal, however the model should be improved.

4.4. Naïve Bayes Method:

Naïve Bayes Classifier is a kind of a ML algorithm, which is based on Bayes' Theorem, for classifying the attributes. The model performance outcome of Naïve Bayes Method is given in Table 6. It has an accuracy rate of 0.73, low p-value, 0.59 sensitivity rate and 0.79 specificity rate. The detailed conditional probabilities and A-priori probabilities table can be seen in Code file.

Table 6. Confusion Matrix and Model Specific Metrics for Naïve Bayes

Confusion Matrix	**Actual**	
Prediction	**1**	**2**
1	247	171
2	170	660

Accuracy	0.7268
95%CI	(0.7011, 0.7513)
No Information Rate	0.6659
P Value	2.06 e-06
Kappa	0.3863
Mcnemar's Test P-Value	1
Sensitivity	0.5923
Specificity	0.7942
Pos Prediction Value	0.5909
Neg Prediction Value	0.7952
Prevalence	0.3341
Detection Rate	0.1979
Detection Prevalence	0.3349
Balanced Accuracy	0.6933