STARTUP PREDICTION ANALYTICAL REPORT

Section 1: Business Objective & Data Structure

Business Objective

Investment strategies for investing in start-up companies are widely based on intuition or past experience. As a result, investors rely primarily on the need being addressed, background of the founders, size of the market being addressed and the ability of the company to scale after tasting early success. The question we pose here is, "can we perform some rigorous analysis that can be used to identify relevant factors and score prospective start- ups based on their potential to be successful". This model/ analysis will then allow investors to make more informed decisions and rely less on their intuitions.

Data Structure

The data has already been split into train, 'CAX_Startup_Train', and test data, 'CAX_Startup_Test'. Data dictionary was also provided that describes all the 51 variables included. The Target variable is a binary class with 234 known observations in the train set and 80 observations to be scored. I combined the data set to wholisticall verify the quality and pre-process all the independent variables for modeling and scoring. I ran a missing values check to and found out there were 80 observations that contains missing values.

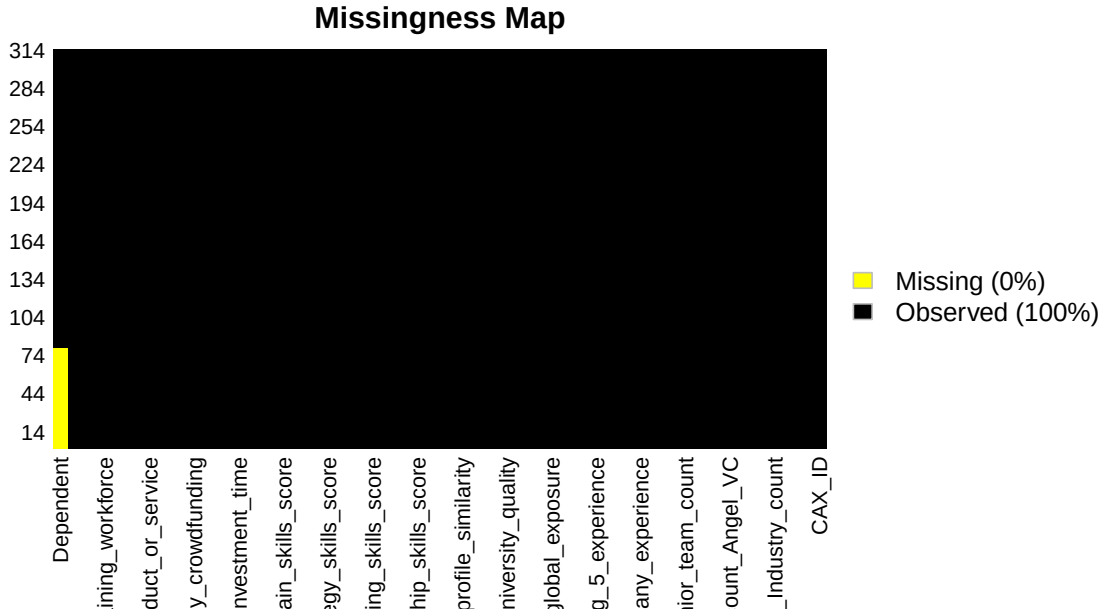


Fig 1 Missing Values

The plot of the missing values showed that they only occurred within the target variable 'Dependent' representing the observations to be predicted. The remainder of the data has no missing value. A glimpse of the data frame showed that numeric and character data types. I converted some of the all of the character data type to categorical data and some coded as integer to categorical data.

Section 2: Exploratory Data Analysis

To effectively explore the data and get insights from the visualizations, I split the data back to train and test and then split the traininto two sub-data frames, continuous and categorical data frames.

I started by plotting the distribution of the continuous variables before moving on the the categorical variables. I grouped some of the plots for comparison and the first set of plots were to understand the distribution of investors to the startups.

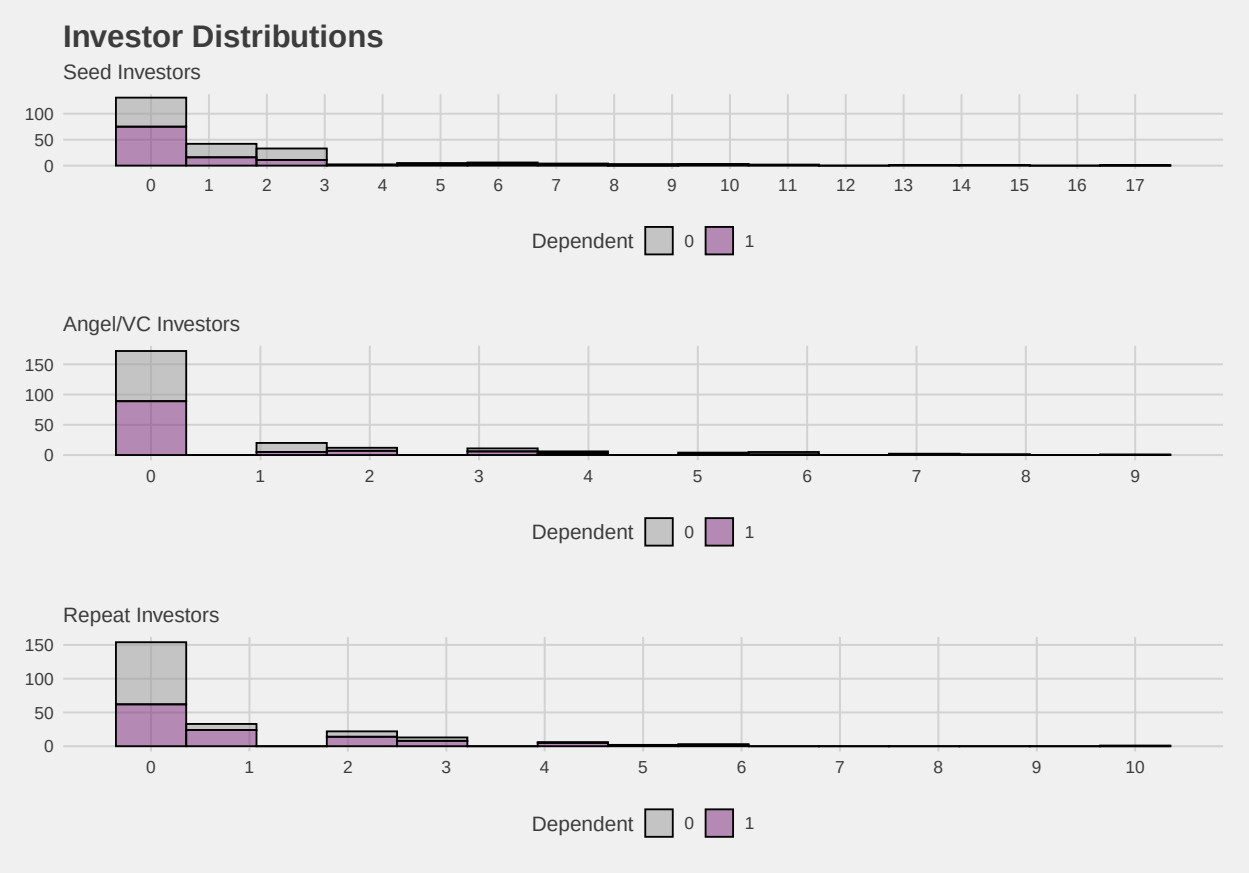


Fig. 2.1 Investor Distributions

We find that all distributions are right skewed with evidence of outliers. In addition, over 75% of the startups had no investors in seed or as Angel/VC. To follow up on this I then check to see how these variables are correlated with each other.

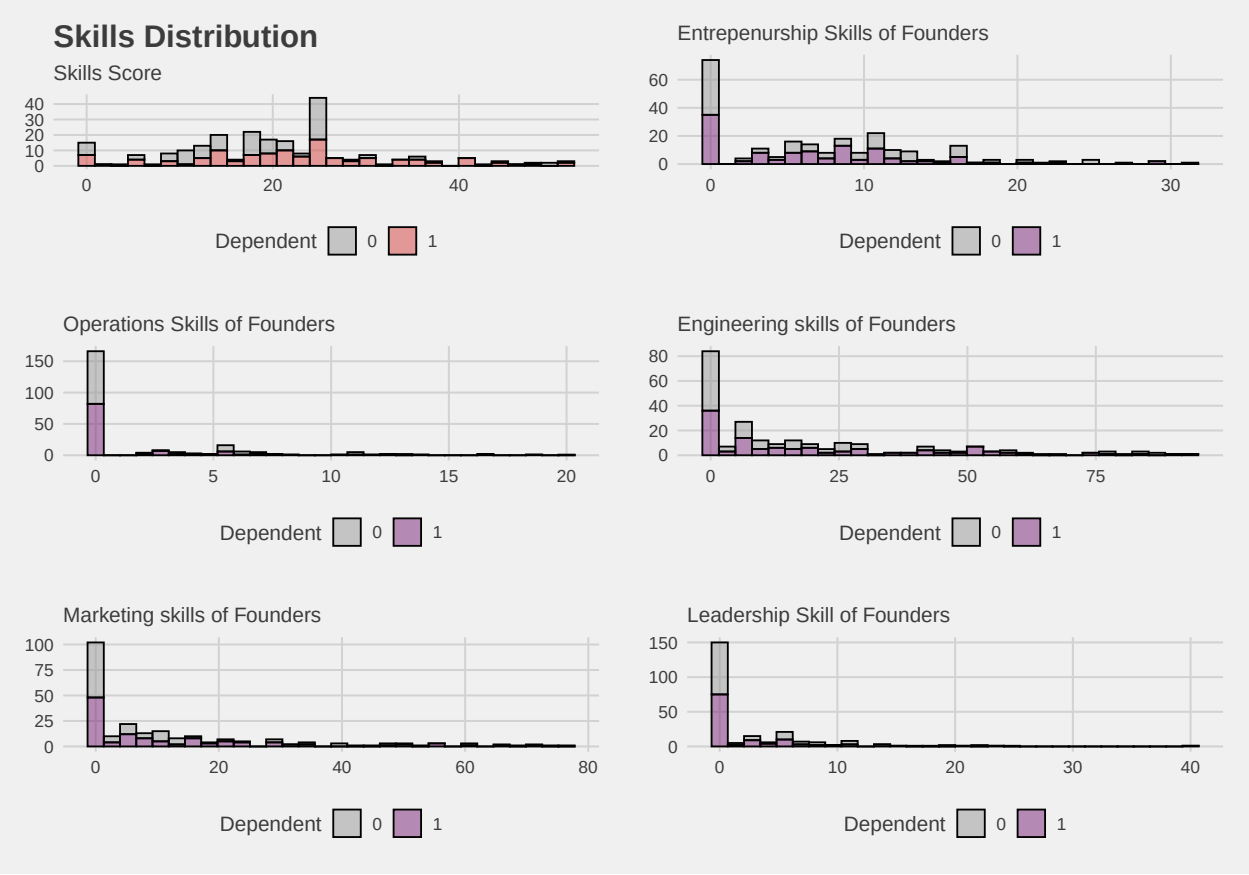
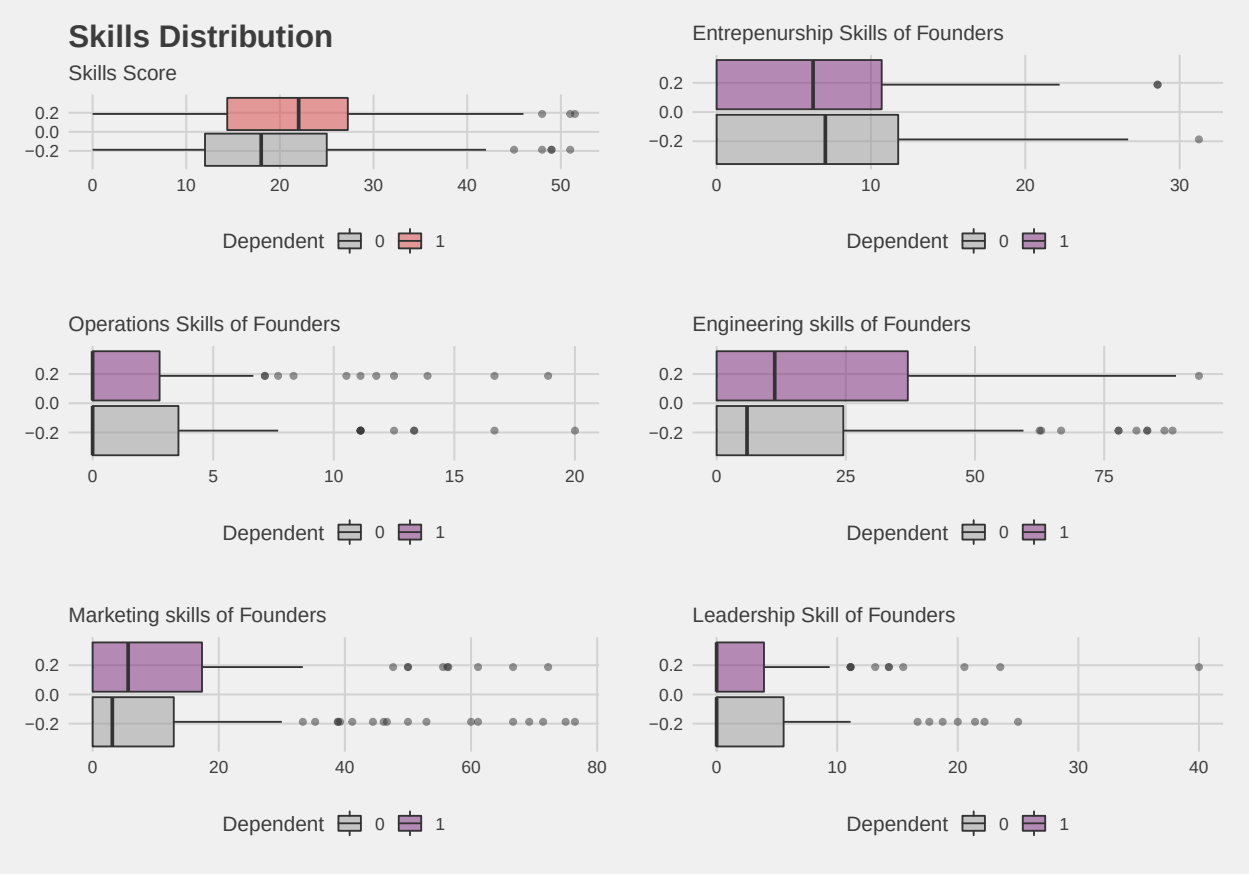

Skills Distribution
Skills Score
Dependent 0 1
Entrepenurship Skills of Founders
Dependent 0 1
Operations Skills of Founders
Dependent 0 1
Engineering skills of Founders
Dependent 0 1
Marketing skills of Founders
Dependent 0 1
Leadership Skill of Founders
Dependent 0 1

Fig 3.2 Skills score

I compared the skills score to all the skill types distribution and found that the looks like the skills score captures has a more normally distributed and all the other skill types are skewed to the right.

Next we check for outliers by making boxplots of the distributions.


Skills Distribution
Skills Score
0.2
0.0
−0.2
0
10
20
30
40
50
Dependent
0
1
Entrepenurship Skills of Founders
0.2
0.0
−0.2
0
10
20
30
Dependent
0
1
Operations Skills of Founders
0.2
0.0
−0.2
0
5
10
15
20
Dependent
0
1
Engineering skills of Founders
0.2
0.0
−0.2
0
25
50
75
Dependent
0
1
Marketing skills of Founders
0.2
0.0
−0.2
0
20
40
60
80
Dependent
0
1
Leadership Skill of Founders
0.2
0.0
−0.2
0
10
20
30
40
Dependent
0
1

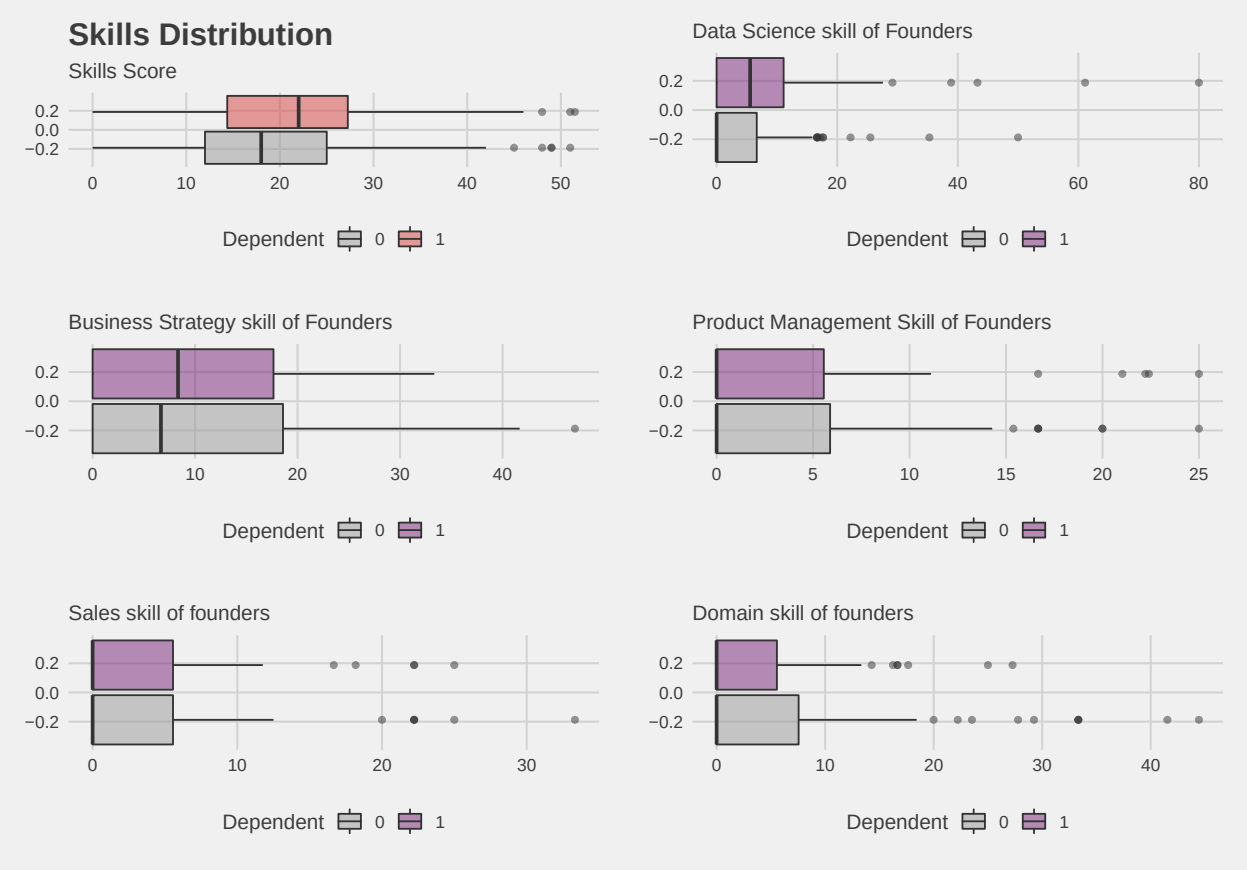


Fig 3.3 Skills set boxplots

The box plots further confirm the distribution of the observation earlier analyzed in the histogram plots. It showed that we have outliers in all of the categories thereby having need to either be log transformed of binning to improve the model we are about to build. Before I make this decision, I would like to see how much information value can be derived from these continuous variables set in categorizing the two groupd of dependent variable.

Section 4: Variable Selection Using Information Value

I used Information Value to understand how well each of the 51 independent variables is able to distinguish the two groups of dependent variable. The 'IV' of variables was used to select variables for modeling.

After computing the information value of all the independent variables, I selected values between 0.1 and 0.5 to build my model. The The selected variables with their IV is as shown in Fig 4.1 below.

```
##                                Variable InformationValue Strength
## 1              Company_avg_investment_time        0.4879221   Strong
## 2                   Company_business_model        0.4272220   Strong
## 3           Founders_Marketing_skills_score        0.2968583   Strong
## 4            Company_repeat_investors_count        0.2899460   Strong
## 5                  Company_analytics_score        0.2297483   Strong
## 6                Company_competitor_count        0.2243885   Strong
## 7              Company_1st_investment_time        0.2112313   Strong
## 8               Founders_Industry_exposure        0.2006278   Strong
## 9              Founders_Sales_skills_score        0.2004713   Strong
```

```
## 10             Founders_Domain_skills_score        0.1999887  Average
## 11       Founders_Data_Science_skills_score        0.1966582  Average
## 12   Founders_Entrepreneurship_skills_score        0.1809043  Average
## 13                    Company_advisors_count        0.1509991  Average
## 14              Company_investor_count_seed        0.1503658  Average
## 15                    Company_crowdfunding        0.1405221  Average
## 16                        Company_big_data        0.1405221  Average
## 17  Founders_Business_Strategy_skills_score        0.1365344  Average
## 18               Founder_university_quality        0.1341797  Average
## 19 Founders_Product_Management_skills_score        0.1334675  Average
## 20                    Founders_publications        0.1294134  Average
## 21        Founders_Engineering_skills_score        0.1215350  Average
## 22                 Founders_global_exposure        0.1082216  Average
```

Fig 4.1 Information Value

Section 5: Model Building

I used a stepwise logistic regression model to select the best combination of variables fitting the training data. I used a significance level cutoff of 20% to develop my final model.

```
## glm(formula = Dependent ~ Company_competitor_count + Company_1st_investment_time +
##     Founders_Data_Science_skills_score + Company_big_data + Founders_publications +
##     Founders_global_exposure, family = binomial(link = logit),
##     data = train_final)
```

```
## # A tibble: 8 x 5
##   term                              estimate std.error statistic p.value
##   <chr>                                <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)                         -0.126     0.438     -0.287 0.774
## 2 Company_competitor_count            -0.0911    0.0534    -1.71  0.0878
## 3 Company_1st_investment_time          0.0272    0.0107     2.54  0.0110
## 4 Founders_Data_Science_skills_score   0.0314    0.0188     1.67  0.0943
## 5 Company_big_dataYes                  1.98      0.839      2.37  0.0180
## 6 Founders_publicationsMany           -0.788     0.544     -1.45  0.148
## 7 Founders_publicationsNone           -1.33      0.448     -2.97  0.00302
## 8 Founders_global_exposureYes          0.912     0.341      2.68  0.00741
```

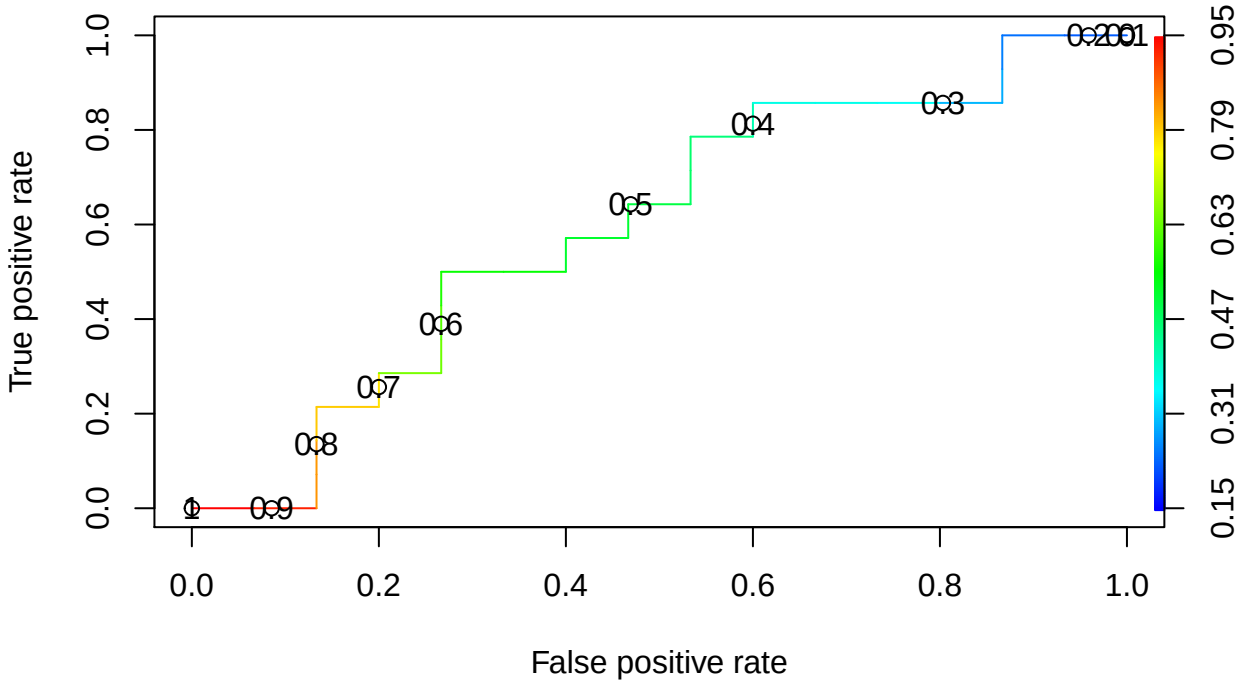
Fig.5.1 Model output

I used 'Hosmer-Lemeshow'test to determine the overall goodness of fit for the regression model with a p-value of 0.05.

The output above shows that the model is fitting the data well with a p-value well above 0.05.

Section 6: Predicting Test Score and Model Evaluation

I used the ROC, AUC and confusion matrix to test the accuracy of the model before prediction.




```
## NULL
```

Fig 6.1 ROC curve of model

From the curve we can see that the optimal probability is at 0.4.

```
##    Probability        FPR        TPR
## 1          Inf 0.00000000 0.00000000
## 2    0.9249893 0.06666667 0.00000000
## 3    0.8358797 0.13333333 0.00000000
## 4    0.8325151 0.13333333 0.07142857
## 5    0.7963323 0.13333333 0.14285714
## 6    0.7920494 0.13333333 0.21428571
## 7    0.7728163 0.20000000 0.21428571
## 8    0.6488985 0.20000000 0.28571429
## 9    0.6398651 0.26666667 0.28571429
## 10   0.6204218 0.26666667 0.35714286
## 11   0.5760387 0.26666667 0.42857143
## 12   0.5572771 0.26666667 0.50000000
## 13   0.5536467 0.33333333 0.50000000
## 14   0.5169798 0.40000000 0.50000000
## 15   0.5166717 0.40000000 0.57142857
## 16   0.5159458 0.46666667 0.57142857
## 17   0.5009368 0.46666667 0.64285714
## 18   0.4766525 0.53333333 0.64285714
## 19   0.4713460 0.53333333 0.71428571
```

```
## 20   0.4582011 0.53333333 0.78571429
## 21   0.4161002 0.60000000 0.78571429
## 22   0.3742400 0.60000000 0.85714286
## 23   0.3700383 0.66666667 0.85714286
## 24   0.3606379 0.73333333 0.85714286
## 25   0.3004480 0.80000000 0.85714286
## 26   0.2911129 0.86666667 0.85714286
## 27   0.2546555 0.86666667 0.92857143
## 28   0.2420103 0.86666667 1.00000000
## 29   0.2285865 0.93333333 1.00000000
## 30   0.1542797 1.00000000 1.00000000
```

Fig 6.2 TPR vs FPR Probability Table

The ROC Curve can be seen for probability point causing largest separation between 'TPR' and 'FPR'. For this we can plot the values in the Table above (Fig 6.2) as given below:

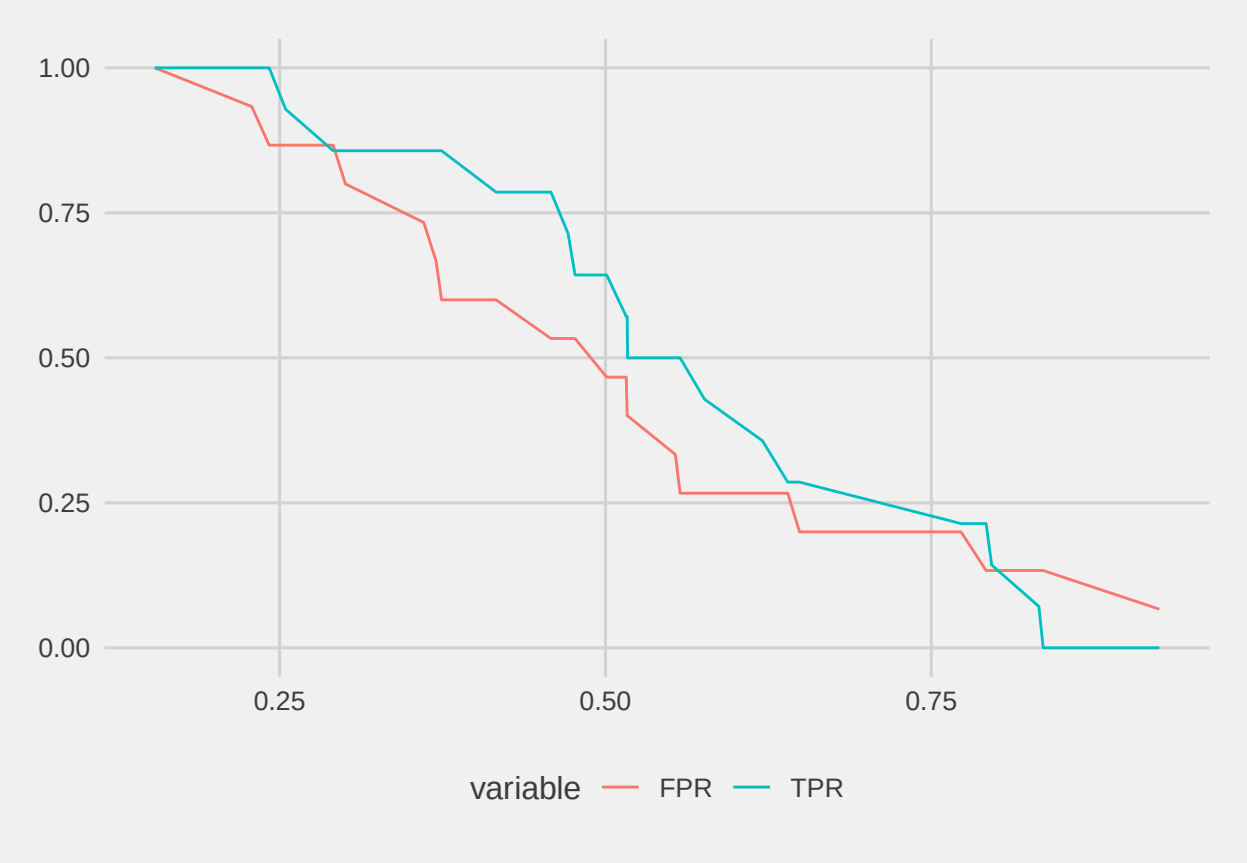


Fig 6.3 Probability Curve

The chart above also shows the optimal probability cut-off of 0.42 which can also be concluded by inspecting the probability table in Fig 6.2 above.Next we use the confusion matrix to see the model accuracy using the probability threshold determined above.

```
##      obs
## pred 0  1
##    0 7  3
##    1 8 11
## attr(,"class")
```

```
## [1] "confusion.matrix"
```

We can now deploy the model to the data we want to score.

```
##    CAX_ID Dependent
## 1       1         1
## 2     249         0
## 3     282         1
## 4       2         1
## 5      24         1
## 6     124         1
## 7     212         1
## 8     224         0
## 9     242         1
## 10    245         0
## 11    246         0
## 12    255         1
## 13    264         1
## 14    267         1
## 15    269         1
## 16    285         0
## 17    273         0
## 18    276         1
## 19    278         0
## 20    286         0
## 21    311         1
## 22    294         0
## 23    296         1
## 24    301         1
## 25    305         0
## 26    308         1
## 27      4         0
## 28    313         0
## 29      6         1
## 30     38         1
## 31     39         0
## 32     21         0
## 33     28         1
## 34     29         0
## 35     32         1
## 36     37         1
## 37     41         1
## 38     49         1
## 39    103         0
## 40     47         1
## 41     48         0
## 42    104         0
## 43     63         0
## 44     65         0
## 45     67         0
## 46     69         0
## 47     74         1
## 48     80         1
## 49    107         1
```

```
## 50     89         1
## 51     95         1
## 52     96         1
## 53    115         1
## 54    126         1
## 55    106         1
## 56    128         1
## 57    132         0
## 58    117         1
## 59    176         0
## 60    180         1
## 61    181         1
## 62    191         1
## 63    149         1
## 64    153         0
## 65    155         0
## 66    160         0
## 67    161         0
## 68    162         1
## 69    166         1
## 70    167         0
## 71    170         1
## 72    172         0
## 73    196         0
## 74    215         0
## 75    232         0
## 76    238         0
## 77    210         1
## 78    216         1
## 79    225         0
## 80    229         0
```