What Factors Affect Credit Scores?

Final Project for Introduction to Data Science

December 7, 2022

Team 2 Thunder

Brooklyn Chen   HaeLee Kim   Sudhanshu Deshpande   Upmanyu Singh

About the Dataset

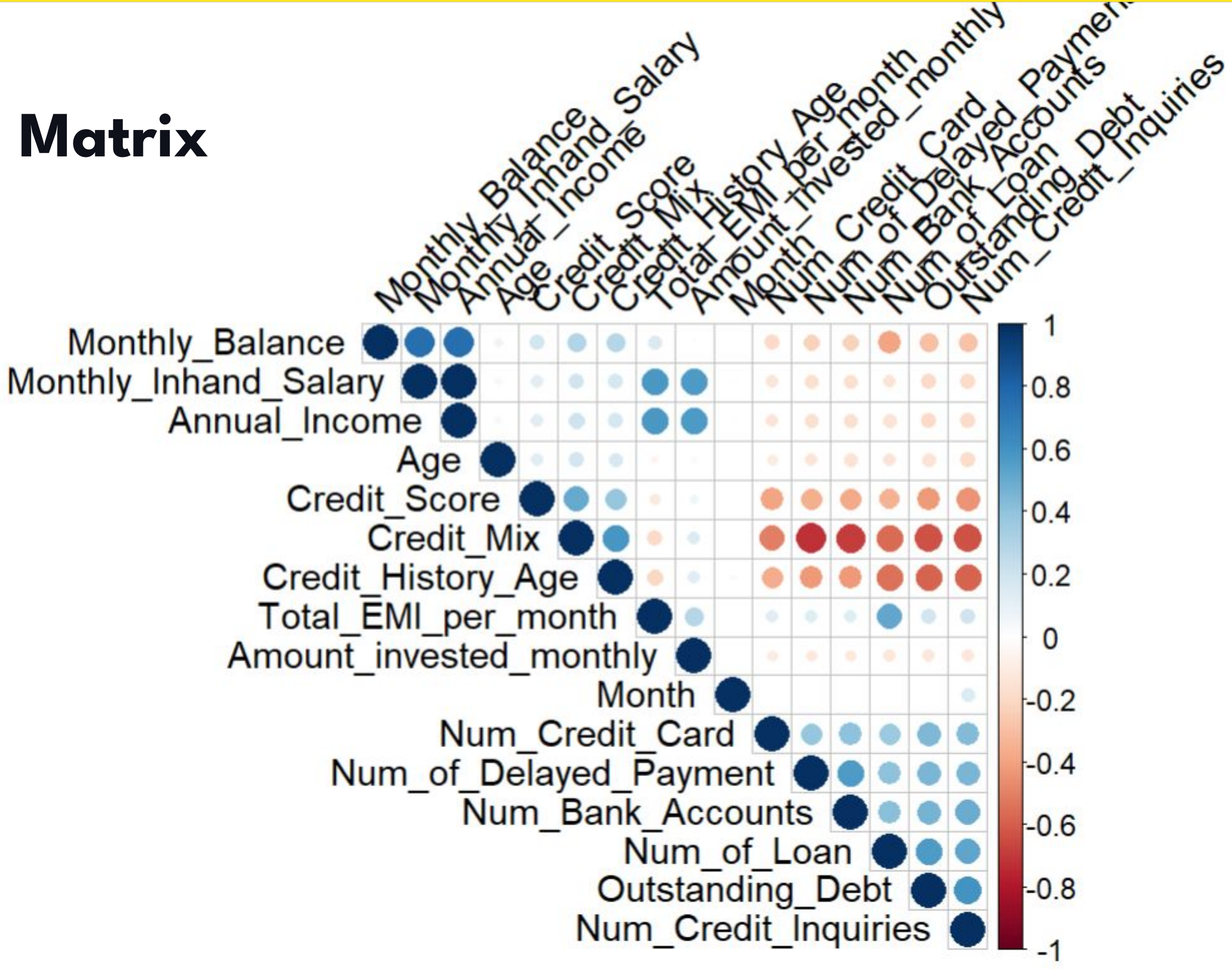
a. Source : Credit Classification in Kaggle
https://www.kaggle.com/datasets/parisrohan/credit-score-classification

b. 30349 Observations of 17 Variables
- Dependent Variable: Credit Score
- Main Independent Variables: Credit Mix, Credit History Age, Monthly Balance
- More Independent Variables: Number of Delayed Payment, Total EMI, Age, Outstanding Debt, Number of Bank Accounts, Number of Credit Card, Number of Loans, Number of Credit Inquiries, Monthly Inhand Salary, Amount Invested Monthly, Annual Income, Month

For Model Analyses

- Dropped nulls and outliers for all variables (used ezids outlierKD2)
- Dropped observations with age below 18 and over 100
- Models used:
  - Linear Regression ○ Logit Regression ○ KNN ○ Decision Tree
- Graphs used:
  - Correlation plot

Correlation Matrix



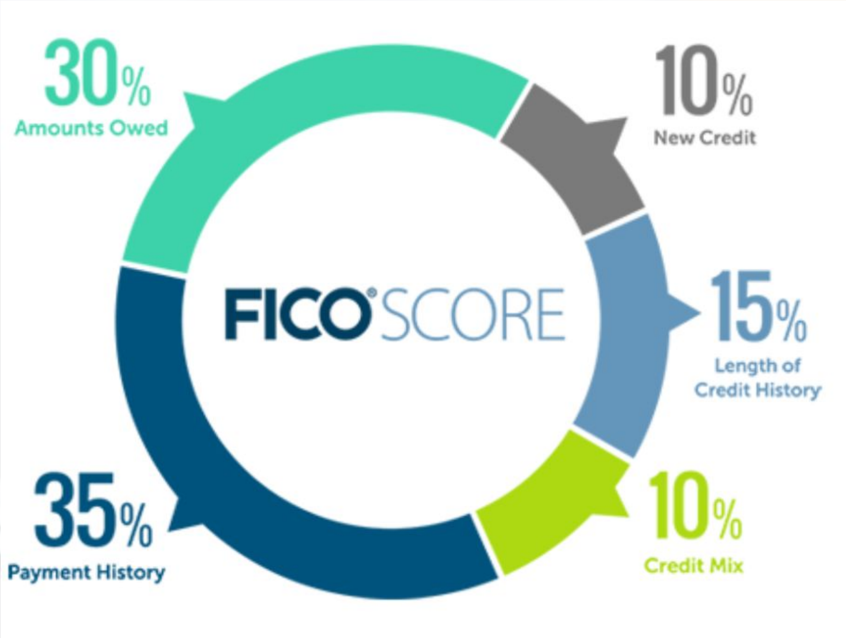
Midterm Summary

a. Credit mix and credit score are not independent of each other. This was done when we tested independence between them using the Chi-square test.

b. Number of delayed payments, Total EMI per month, and Age have undergone the ANOVA test for the difference between the mean of the three groups. All the variables have significant mean differences between the groups - poor, standard, and good credit score groups. Furthermore, we verified the ANOVA test by doing the post hoc Tuckey test, which also tells us that all the groups are significantly different from each other.

Linear Regression Analysis

How is the Credit Score Calculated?



SMART Question 1

Five independent variables - payment history, amounts owed, length of credit history, new credit, and credit mix - are known to affect our credit score.

Do these (above listed) five independent variables affect the credit score?

Note: Although variables used to calculate the credit score/fico score are **five**, we have only **three** out of five variables so will use three for our analysis.

SMART Analysis 1 - Linear Regression

```
lm(formula = Credit_Score ~ Monthly_Balance + Credit_History_Age +
    Credit_Mix, data = dfscale)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1801 -0.5229  0.1343  0.4475  2.6669

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.662e-16  4.911e-03   0.000        1
Monthly_Balance    2.791e-02  5.193e-03   5.374 7.74e-08 ***
Credit_History_Age 1.290e-01  6.113e-03  21.100  < 2e-16 ***
Credit_Mix         4.220e-01  6.141e-03  68.714  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8555 on 30345 degrees of freedom
Multiple R-squared:  0.2681,    Adjusted R-squared:  0.2681
F-statistic:  3706 on 3 and 30345 DF,  p-value: < 2.2e-16
```

SMART Answer 1

Do variables - amounts owed, length of credit history, and credit mix - affect the credit score?

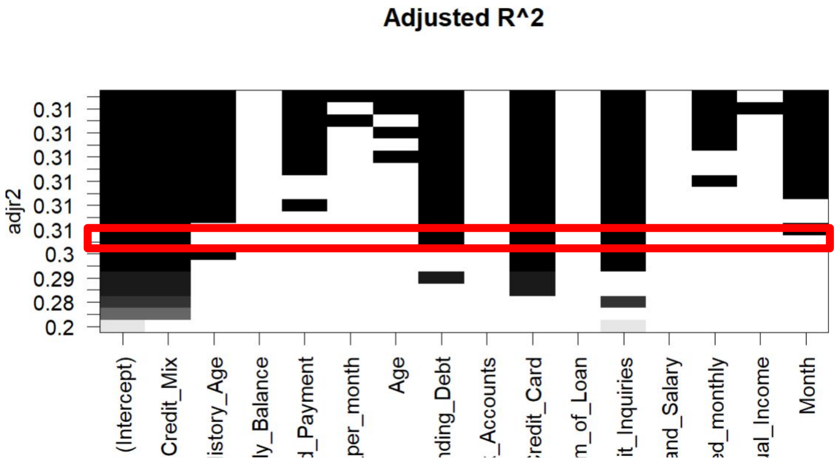
Yes, we can find that monthly balance, credit history age and credit mix significantly increase the credit score.

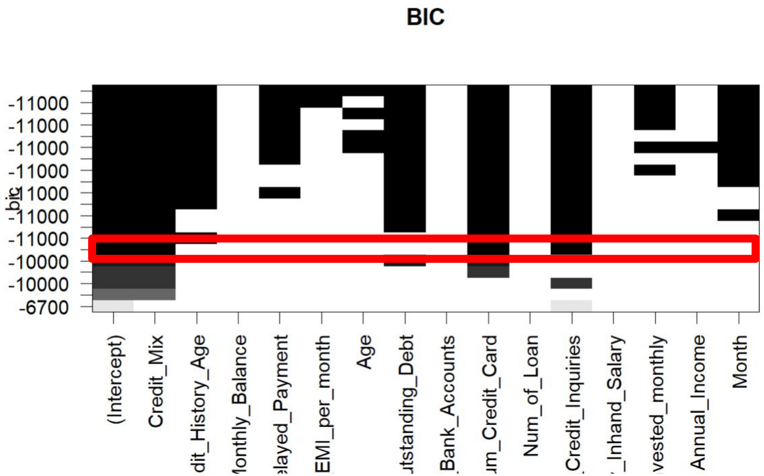
SMART Question 2

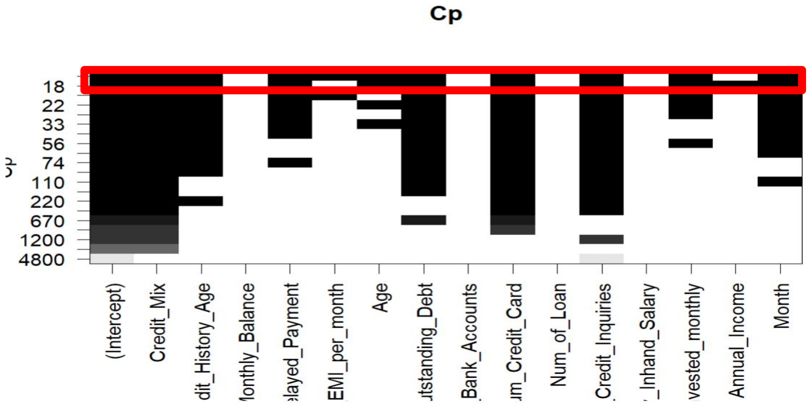
We include more variables including the variables used in previous analysis – number of credit cards, bank accounts, loans, credit inquiries, and monthly in-hand salary – in our model.

Which model is considered the best fit model?

SMART Analysis 2 - Feature Selection







SMART Analysis 2 - Linear Regression

```
lm(formula = Credit_Score ~ Credit_Mix + Num_Credit_Card + Num_Credit_Inquiries,
    data = dfscale)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1890 -0.5136  0.0430  0.5400  2.8629 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -4.971e-16  4.799e-03    0.00        1    
Credit_Mix            3.104e-01  6.574e-03   47.22   <2e-16 ***
Num_Credit_Card      -1.642e-01  5.642e-03  -29.11   <2e-16 ***
Num_Credit_Inquiries -1.771e-01  6.297e-03  -28.13   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.836 on 30345 degrees of freedom
Multiple R-squared:  0.3011,	Adjusted R-squared:  0.3011 
F-statistic:  4359 on 3 and 30345 DF,  p-value: < 2.2e-16
```

Note: We ran three different models each one have greater R2, lower bic and cp respectively. All three models had the same residual error and adjusted r2. We chose the model with the least BIC because it contains least number of independent variables.

SMART Answer 2

After the feature selection process, the final model's variables are extracted from the BIC result.

credit mix $\xrightarrow{+}$ credit score

number of credit cards $\xrightarrow{-}$ credit score

number of credit inquiries $\xrightarrow{-}$ credit score

"Don't ask your credit score too much."

Logit Regression Analysis

SMART Question 3

From the previous analysis in *Question 1* we found that :
Three variables - amounts owed, length of credit history, and credit mix - are related to credit score in Linear Regression.

Can we verify that 3 variables -listed above- will increase/decrease the chance of being in the good-standard credit score group?

Note: Question 1 and Question 3 are exactly the same. The only difference is we have used different techniques for the analysis. For question 1 we have used linear regression and question 3 logit regression for the analysis

SMART Analysis 3

How to operationalize the dummy variable for y?

	Model 1	**Model 2**	**Model 3**
y = 1	Good	Standard - Good	Good
y = 0	Poor - Standard	Poor	Poor
Conclusion	Overfitting	Good to use	Good to use

SMART Analysis 3

Model 2: 0 for poor & 1 for standard-good

```
glm(formula = Credit_Score2 ~ Monthly_Balance + Credit_History_Age +
    Credit_Mix, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4831  -0.9888   0.5496   0.7831   1.6886

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.0206207  0.0511335  -39.52   <2e-16 ***
Monthly_Balance     0.0011520  0.0001122   10.27   <2e-16 ***
Credit_History_Age  0.0585369  0.0022039   26.56   <2e-16 ***
Credit_Mix          0.7355022  0.0254405   28.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 36432  on 30348  degrees of freedom
Residual deviance: 31761  on 30345  degrees of freedom
AIC: 31769

Number of Fisher Scoring iterations: 4
```

Confusion matrix from Logit Model2

	Predicted 0	Predicted 1	Total
Actual 0	2941	5795	8736
Actual 1	1802	19811	21613
Total	4743	25606	30349

Model 3: 0 for poor & 1 for good credit score

```
glm(formula = Credit_Score3 ~ Monthly_Balance + Credit_History_Age +
    Credit_Mix, family = "binomial", data = subset_logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2198  -0.3369  -0.3092  -0.0743   3.4705

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -8.8043172  0.1220271 -72.151  < 2e-16 ***
Monthly_Balance    -0.0001199  0.0001258  -0.954     0.34
Credit_History_Age  0.0152842  0.0031068   4.920 8.67e-07 ***
Credit_Mix          2.8056585  0.0437135  64.183  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27791  on 30348  degrees of freedom
Residual deviance: 18953  on 30345  degrees of freedom
AIC: 18961

Number of Fisher Scoring iterations: 6
```

Confusion matrix from Logit Model3

	Predicted 0	Predicted 1	Total
Actual 0	23896	1256	25152
Actual 1	3977	1220	5197
Total	27873	2476	30349

SMART Answer 3

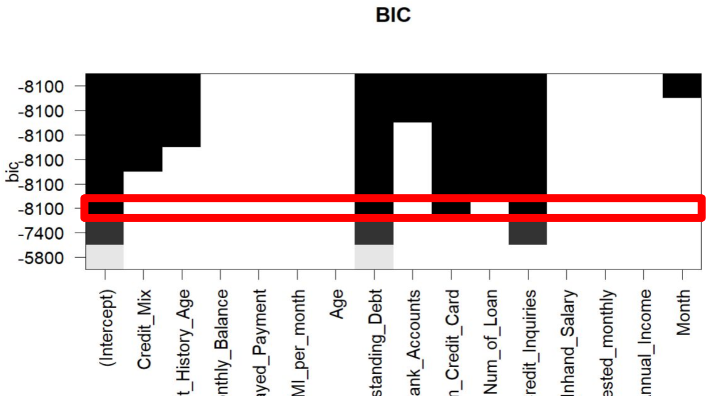
Yes, the more amounts owed, the longer length of credit history, and the good credit mix more likely to be in a good-standard group.

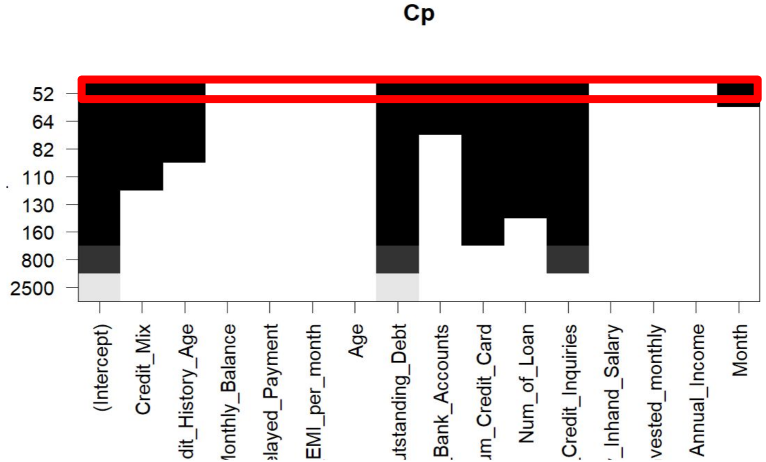
SMART Question 4

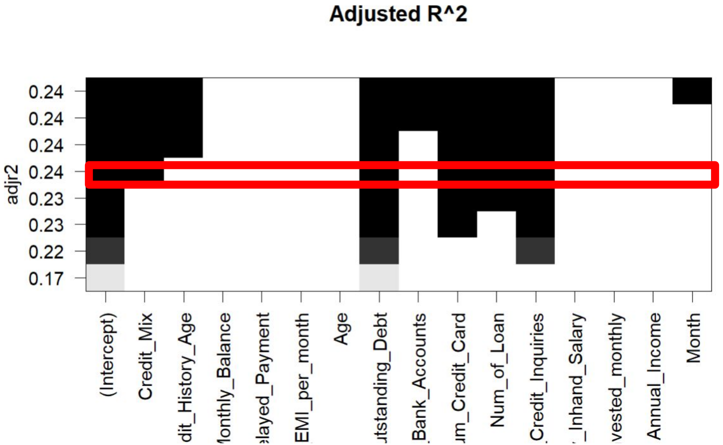
Can we predict the probability of good credit score with sample information by using the best-fit model?

For example: If we know the information of Sudhanshu and Upmanyu, can we predict the probability of getting a higher credit score?

SMART Analysis 4 - Feature Selection







SMART Analysis 4 - Logit Regression

```
glm(formula = Credit_Score2 ~ Outstanding_Debt + Num_Credit_Card +
    Num_Credit_Inquiries, family = "binomial", data = subset2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5109  -0.8229   0.4898   0.6641   2.1989

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           3.671e+00  4.848e-02   75.72   <2e-16 ***
Outstanding_Debt     -5.493e-04  2.039e-05  -26.95   <2e-16 ***
Num_Credit_Card      -1.863e-01  8.359e-03  -22.28   <2e-16 ***
Num_Credit_Inquiries -1.519e-01  4.931e-03  -30.81   <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 36432  on 30348  degrees of freedom
Residual deviance: 29204  on 30345  degrees of freedom
AIC: 29212
```

Confusion matrix from Logit Model4

	Predicted 0	Predicted 1	Total
Actual 0	4345	4391	8736
Actual 1	2257	19356	21613
Total	6602	23747	30349

Note: We ran three different models as per slide 16. All three models have the same confusion matrix. For model 2, the p-value for variable monthly balance is less than 0.05 but for the other two model it p-value exceeds 0.05.

SMART Question 4

	 Sudhanshu	 Upmanyu
Outstanding Debt	800	500
Number of Credit Cards	6	8
Number of Credit Inquiries	2	10

SMART Answer 4

85.9% chance to be in a standard-good score group

vs

59.5% chance to be in a standard-good score group

"Upmanyu, don't ask about your credit score often."

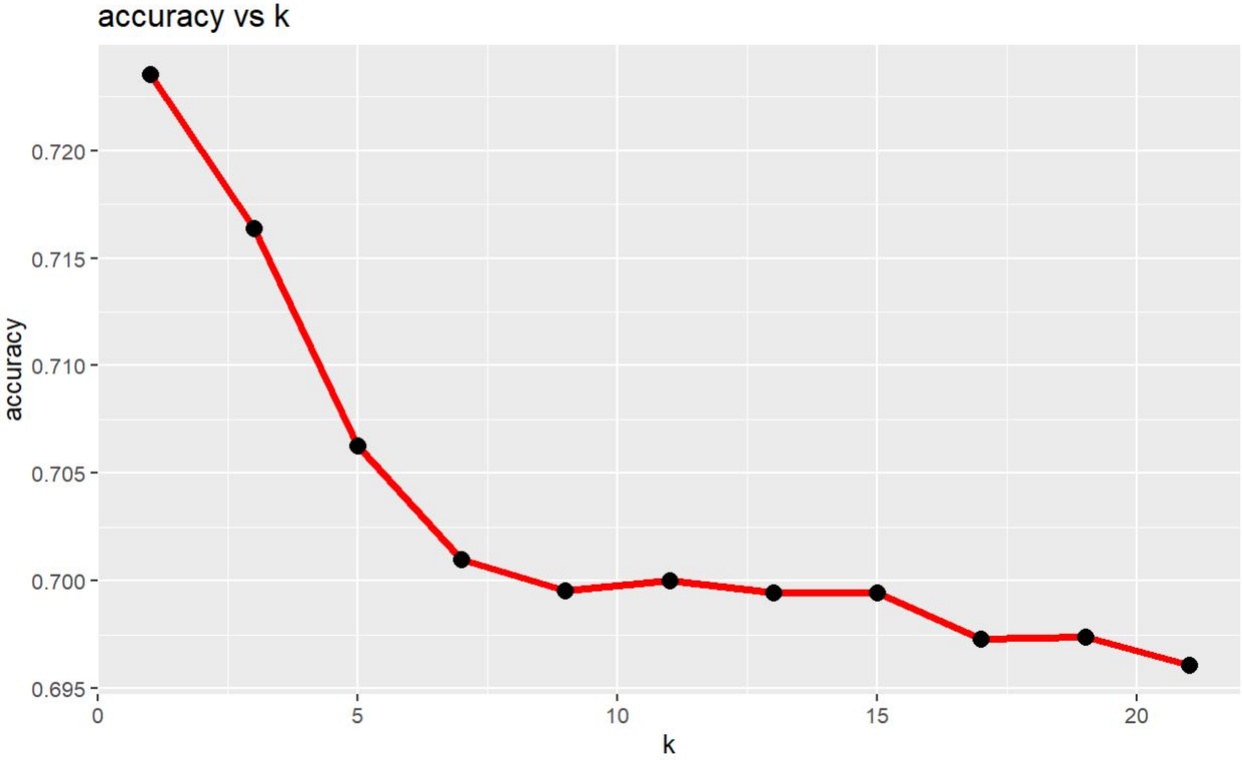
KNN Analysis

SMART Question 5

What is optimal number of n that can be grouped out of the data so that KNN accuracy is highest?

As our dependent variable has 3 groups - poor, standard, and good, we will refer to accuracy as per the model evaluation.

SMART Answer 5



We plotted graph for different accuracy versus k and find out the optimal k.

K=7 is our answer for question 5.

SMART Analysis 5

Confusion Matrix of Knn (K=7)

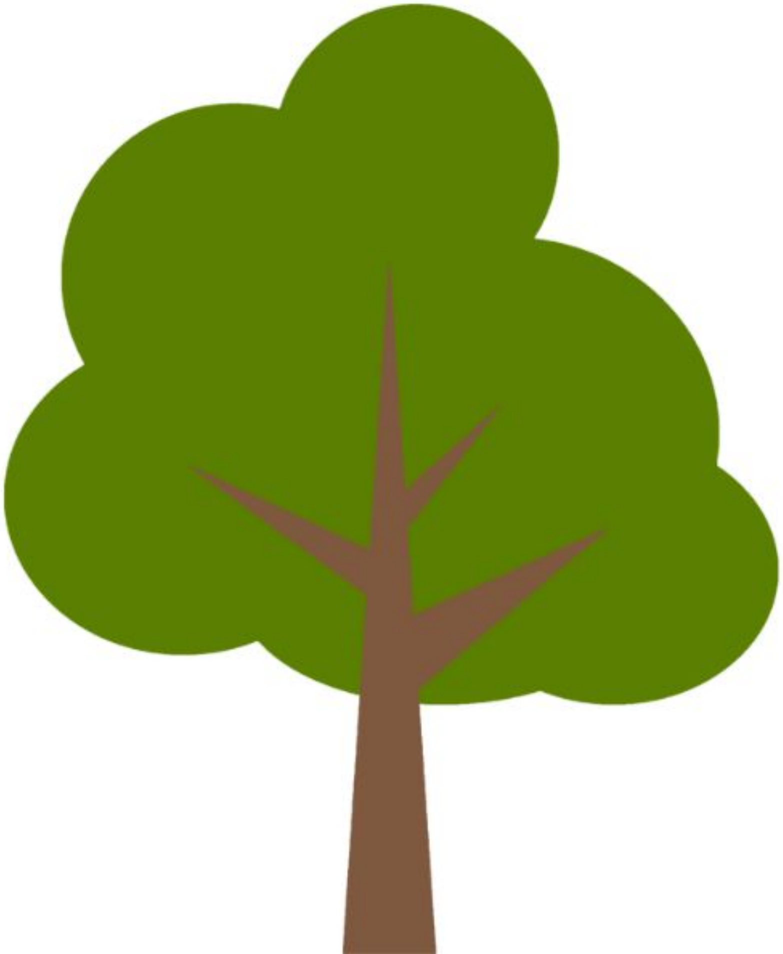
LABELS		Actual		
		1	2	3
Prediction	1	1699	732	87
	2	717	3814	574
	3	204	448	945

Accuracy = 70% on test dataset

Decision Tree Analysis

SMART Question 6

Would Sudhanshu and Upmanyu’s credit score result be the same by tree model?



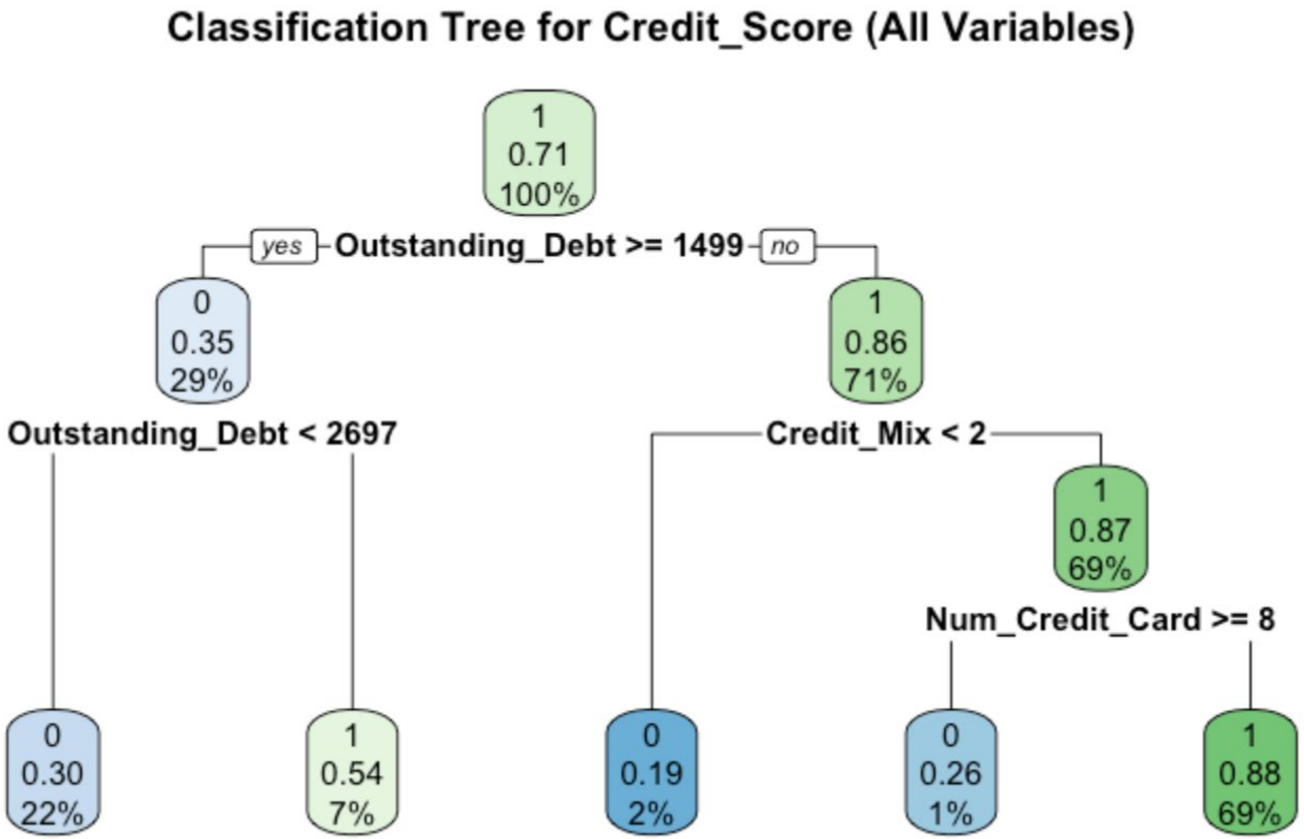
SMART Analysis 6 - Decision Tree

- **Supervised Learning Algorithm**
- **Classification Tree:** Categorical Dependent Variable
- **Regression Tree:** Quantitative Dependent Variable

SMART Analysis 6 - Decision Tree

```
Classification tree:
rpart(formula = Credit_Score2 ~ ., data = subset2, method = "class",
    control = list(maxdepth = 4))

Variables actually used in tree construction:
[1] Credit_Mix        Num_Credit_Card  Outstanding_Debt

Root node error: 8736/30349 = 0.28785

n= 30349

        CP nsplit rel error  xerror      xstd
1 0.288233      0   1.00000 1.00000 0.0090288
2 0.043727      1   0.71177 0.71326 0.0080550
3 0.018773      2   0.66804 0.66976 0.0078668
4 0.010989      3   0.64927 0.65167 0.0077848
5 0.010000      4   0.63828 0.64732 0.0077647
Call:
rpart(formula = Credit_Score2 ~ ., data = subset2, method = "class",
    control = list(maxdepth = 4))
  n= 30349
```

SMART Analysis 6 - Decision Tree



Would Sudhanshu's credit score result be the same by tree model?

Sudhanshu's Profile

Outstanding Debt: 500

Credit Mix: 3

Number of Credit Card: 6

Good Credit Score ?

Bad Credit Score ?

Would Upmanyu's credit score result be the same by tree model?

Upmanyu's Profile

Outstanding Debt: 800

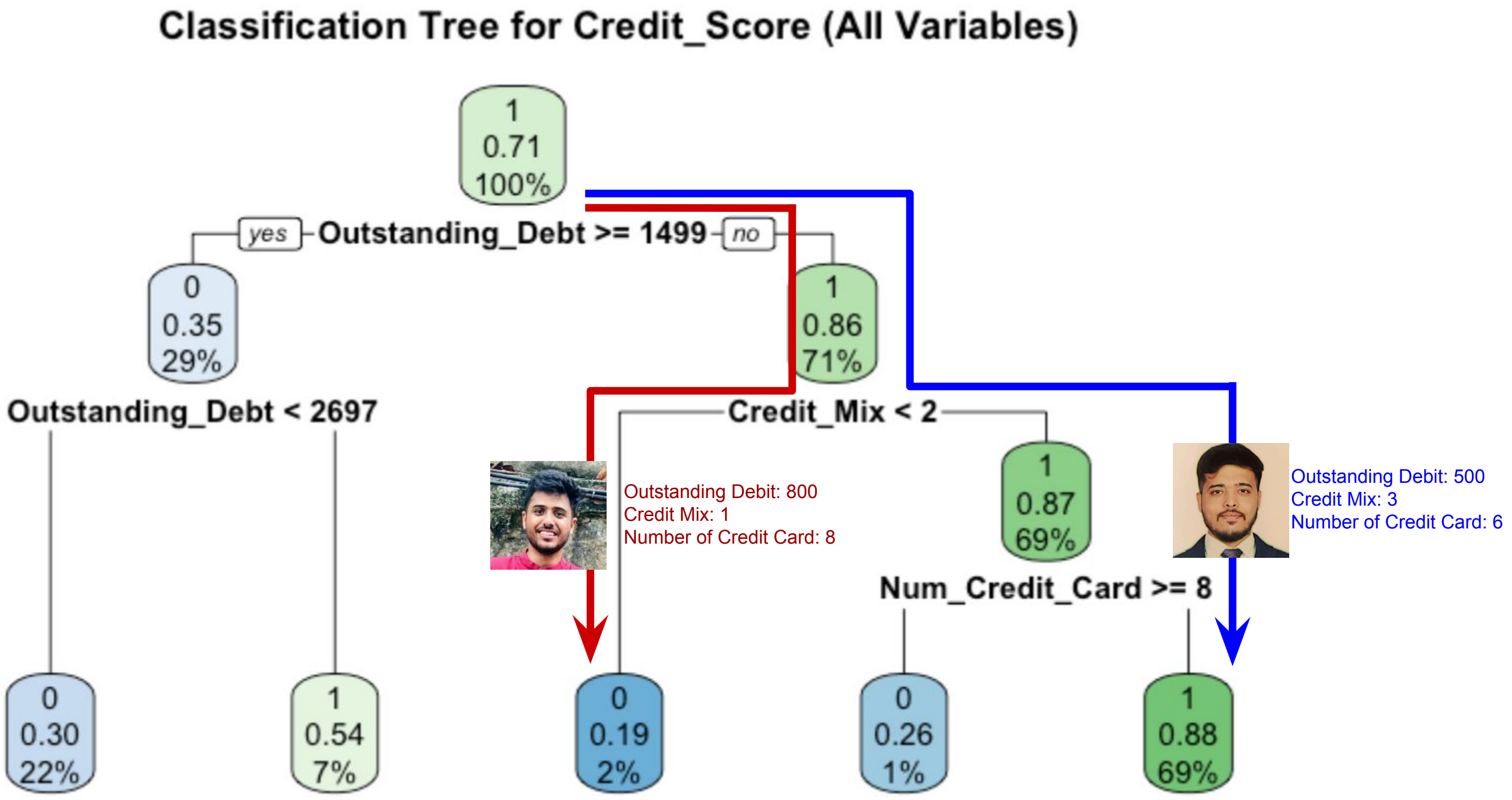
Credit Mix: 1

Number of Credit Card: 8

Good Credit Score ?

Bad Credit Score ?

SMART Answer 6

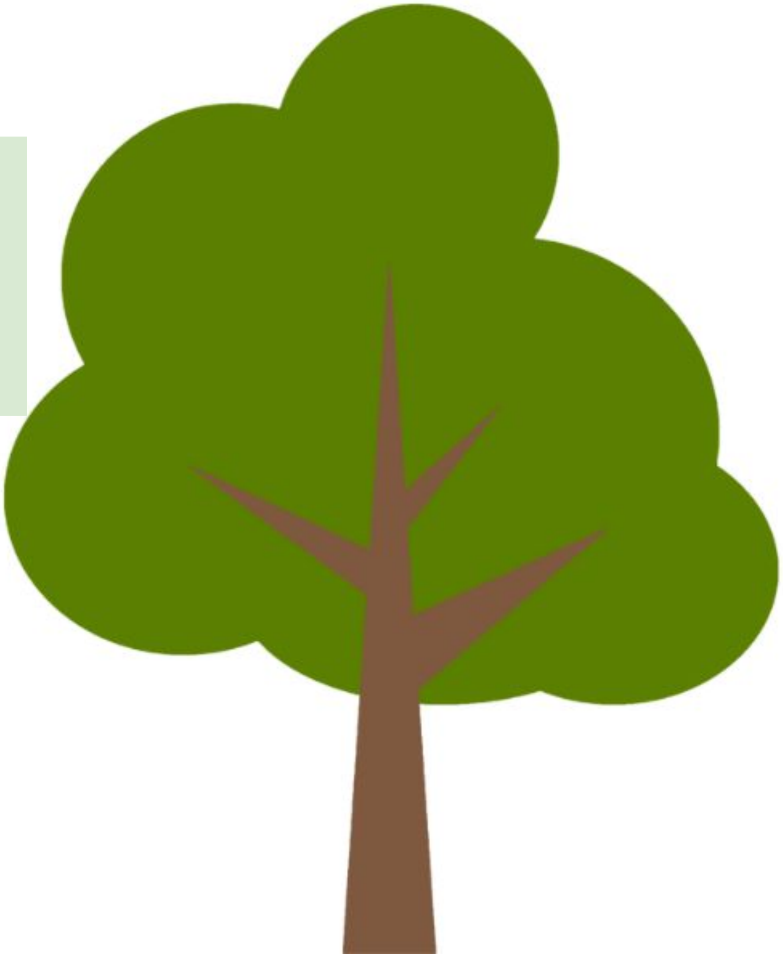


SMART Answer 6

Would Sudhanshu and Upmanyu's credit score result be the same by tree model?

Yes!!

Sudhanshu: Good/Standard
Upmanyu: Bad :(



Summary

1. 3 categories FICO suggested - amounts owed, length of credit history, and credit mix - significantly affect the credit score.
2. The credit mix increases credit scores, but the number of credit cards and credit inquiries decreases credit scores.
3. The more amounts owed, the longer length of credit history, and the higher credit mix means more likely to be in the good-standard group.
4. Don't ask about your credit score often.
5. The results of the decision tree are the same as the prediction by logic regression.

Thank You!