PREDICTION REPORT

AUTHOR
Luke Stucky

PUBLISHED
December 11, 2024

BUSINESS UNDERSTANDING

Following the classification model aimed at reducing credit defaults, the financial institution is now focused on predicting credit scores more accurately to enhance their decision-making process. The classification model helped the institution identify high-risk customers, but predicting credit scores will allow the institution to fine-tune its offerings based on a customer's exact creditworthiness. Accurate credit score predictions can allow the institution to offer personalized financial products, better interest rates, and improve overall customer satisfaction.

As the head data analyst, Luke Stucky now seeks to improve the institution's ability to predict customers' credit scores based on their financial behaviors. This predictive model will provide an understanding of the financial health of applicants, allowing for more informed lending decisions and more targeted risk management strategies.

By enhancing the current credit score models, Luke Stucky aims to answer the following key research questions:

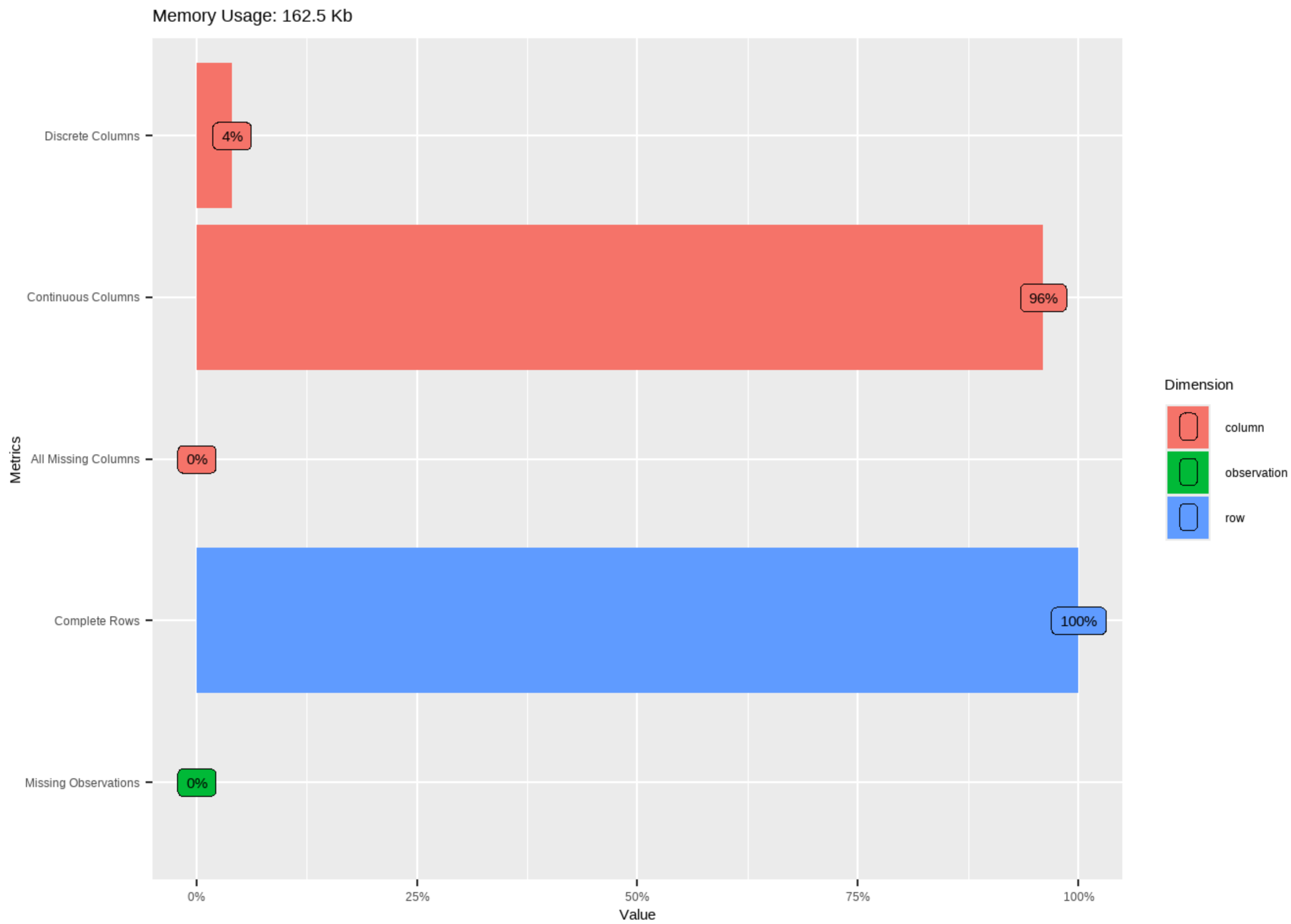
How can financial behaviors such as income, savings, and debt predict an individual's credit score more accurately?

Which factors have the greatest impact on predicting credit scores, and how can we leverage these insights for better decision-making?

This predictive model will provide a deeper understanding of customers' financial profiles, allowing the institution to better tailor its lending and financial products, ensure more responsible lending, and ultimately increase profitability while reducing risk.

DATA UNDERSTANDING

EDA Inspect dataset for Missing Values and Outliers



variables	outliers_cnt	outliers_ratio	outliers_mean	with_mean	without_mean
default	0	0.0	NaN	0.2840000	0.2840000
income	25	2.5	464058.8400000	121610.0190000	112829.2800000
savings	28	2.8	1965481.7857143	413189.5970000	368473.3611111
debt	44	4.4	3736178.0000000	790718.0450000	655152.9424686
r_savings_income	21	2.1	14.9746524	4.0634772	3.8294275
r_debt_income	35	3.5	25.9302571	6.0684492	5.3480727
r_clothing_income	53	5.3	0.1630698	0.0555572	0.0495401
r_education_income	108	10.8	0.2163861	0.0386945	0.0171803
r_entertainment_income	50	5.0	0.6032900	0.1675136	0.1445780
r_fines_income	99	9.9	0.0029394	0.0002910	0.0000000
r_gambling_income	160	16.0	0.0841012	0.0184709	0.0059699
r_groceries_income	37	3.7	0.4151811	0.1564751	0.1465352
r_health_income	66	6.6	0.1741470	0.0523004	0.0436903
r_housing_income	0	0.0	NaN	0.0926080	0.0926080
r_tax_income	3	0.3	0.0869333	0.0250889	0.0249028
r_travel_income	28	2.8	0.8977893	0.2828336	0.2651188

variables	outliers_cnt	outliers_ratio	outliers_mean	with_mean	without_mean
r_utilities_income	17	1.7	0.1330471	0.0546550	0.0532993
r_expenditure_income	26	2.6	1.5795462	0.9436065	0.9266307
cat_debt	56	5.6	0.0000000	0.9440000	1.0000000
cat_credit_card	236	23.6	1.0000000	0.2360000	0.0000000
cat_mortgage	173	17.3	1.0000000	0.1730000	0.0000000
cat_savings_account	7	0.7	0.0000000	0.9930000	1.0000000
cat_dependents	150	15.0	1.0000000	0.1500000	0.0000000
credit_score	34	3.4	402.1470588	586.7120000	593.2080745

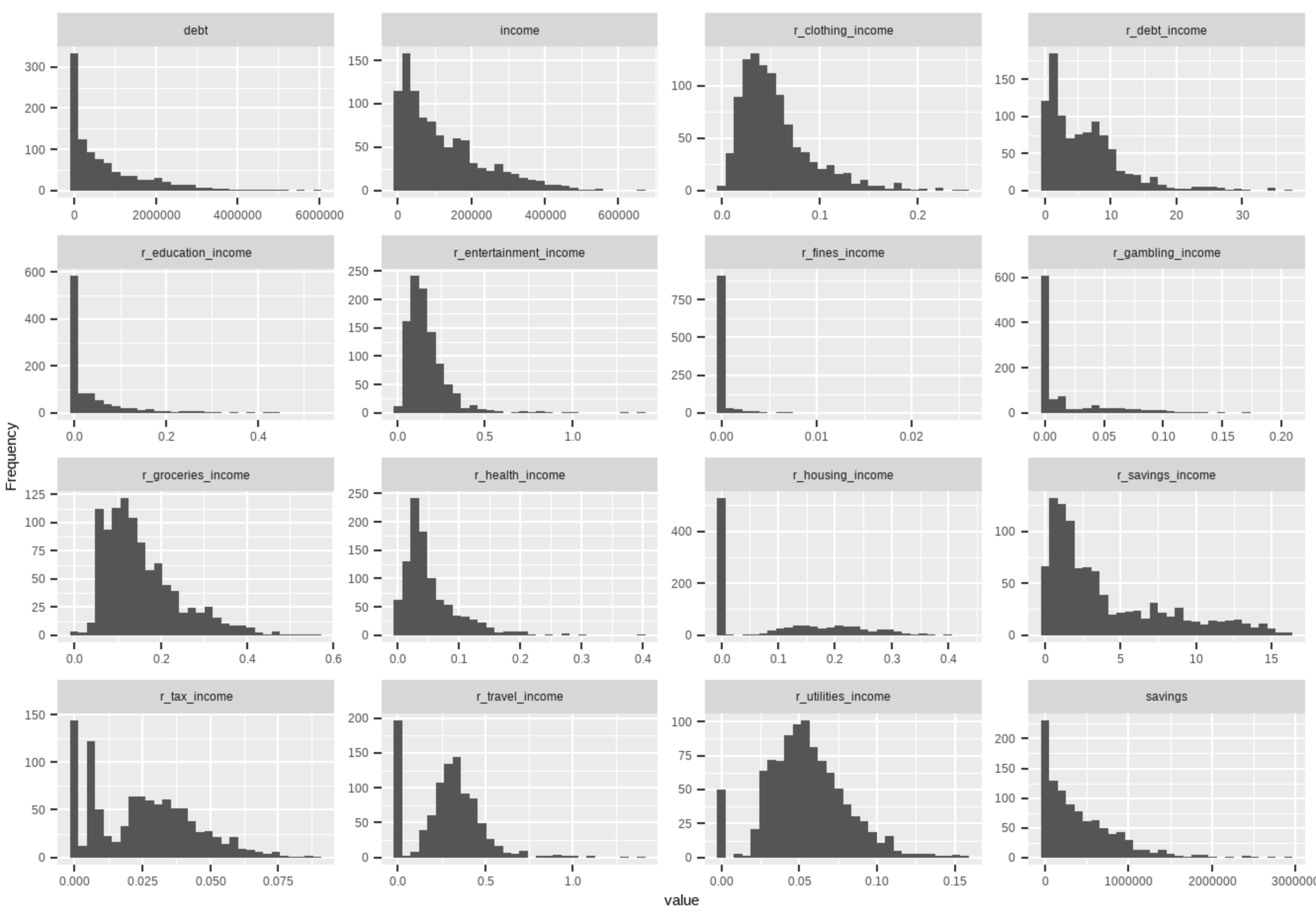
We do not have any missing values in this data. Outliers are important to keep in this data as they can be indicaters of a customer’s credit score.

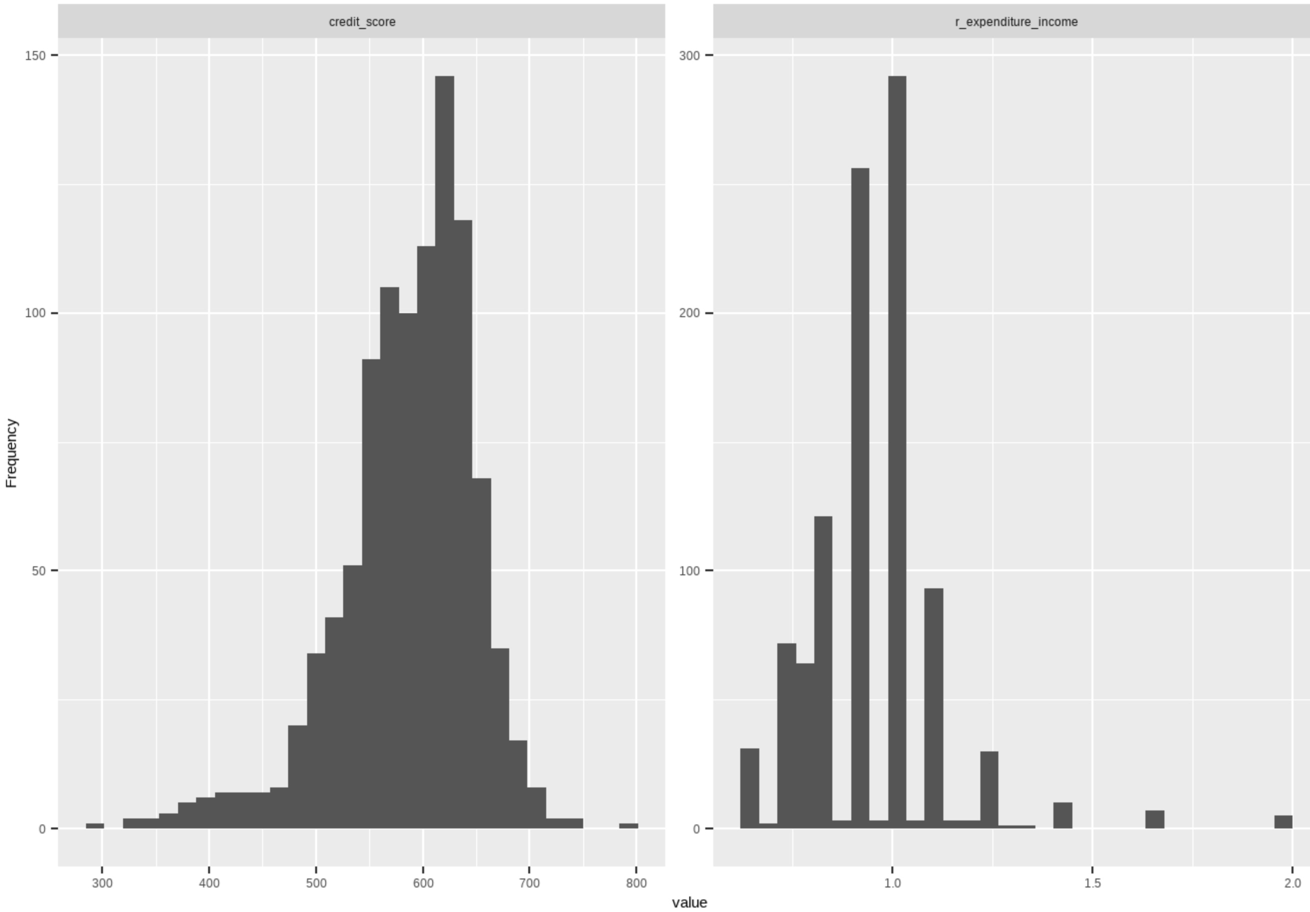
Check default proportion for balance

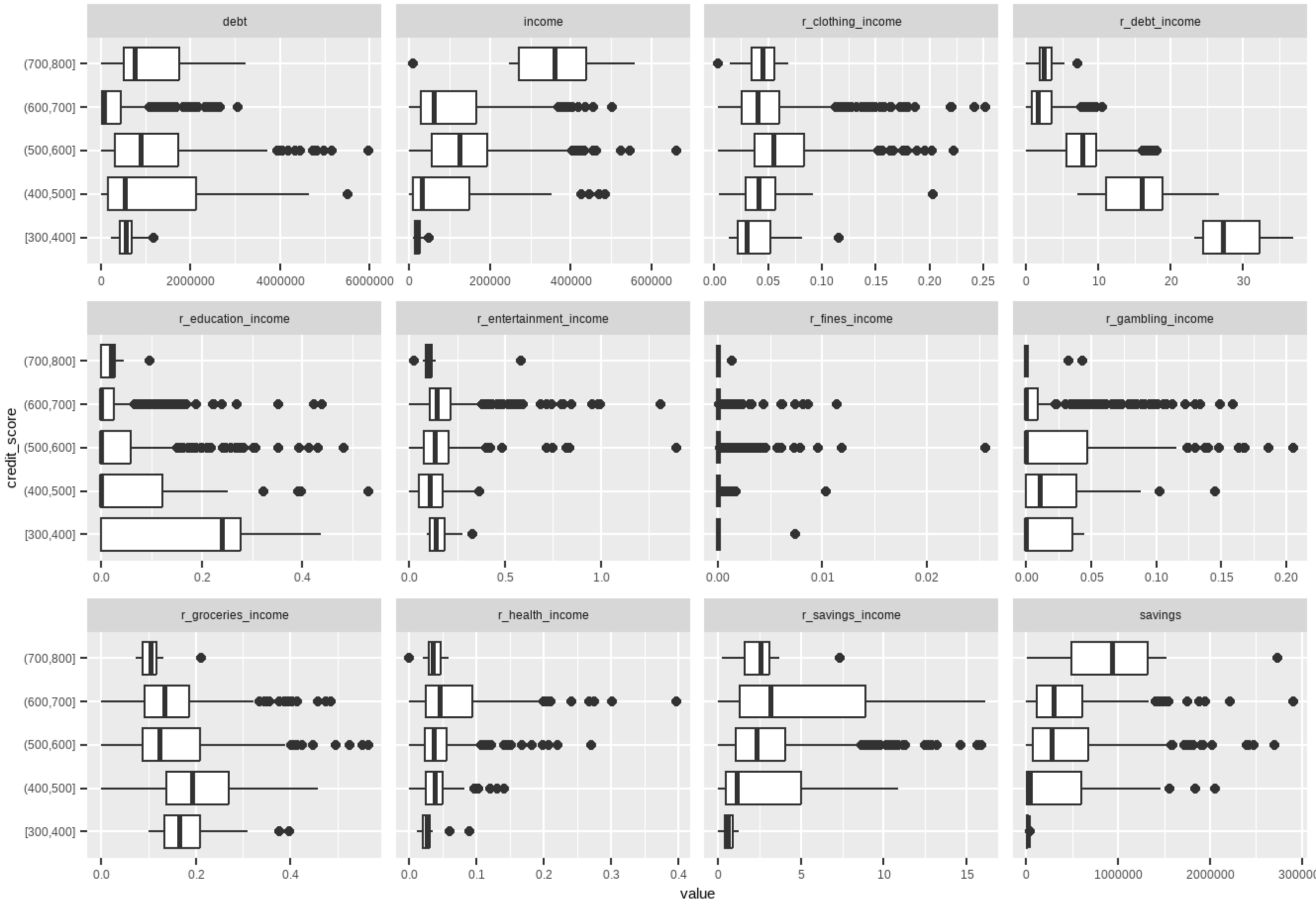
```
                       vars    n      mean        sd    min        max
default                   1 1000      0.28      0.45   0.00       1.00
income                    2 1000 121610.02 113716.70   0.00  662094.00
savings                   3 1000 413189.60 442916.04   0.00 2911863.00
debt                      4 1000 790718.04 981790.39   0.00 5968620.00
r_savings_income          5 1000      4.06      3.97   0.00      16.11
r_debt_income             6 1000      6.07      5.85   0.00      37.00
r_clothing_income         7 1000      0.06      0.04   0.00       0.25
r_education_income        8 1000      0.04      0.07   0.00       0.53
r_entertainment_income    9 1000      0.17      0.14   0.00       1.40
r_fines_income           10 1000      0.00      0.00   0.00       0.03
r_gambling_income        11 1000      0.02      0.03   0.00       0.21
r_groceries_income       12 1000      0.16      0.09   0.00       0.56
r_health_income          13 1000      0.05      0.05   0.00       0.40
r_housing_income         14 1000      0.09      0.11   0.00       0.43
r_tax_income             15 1000      0.03      0.02   0.00       0.09
r_travel_income          16 1000      0.28      0.20   0.00       1.40
r_utilities_income       17 1000      0.05      0.03   0.00       0.16
r_expenditure_income     18 1000      0.94      0.17   0.67       2.00
cat_gambling             19 1000       NaN        NA    Inf       -Inf
cat_debt                 20 1000      0.94      0.23   0.00       1.00
cat_credit_card          21 1000      0.24      0.42   0.00       1.00
cat_mortgage             22 1000      0.17      0.38   0.00       1.00
cat_savings_account      23 1000      0.99      0.08   0.00       1.00
cat_dependents           24 1000      0.15      0.36   0.00       1.00
credit_score             25 1000    586.71     63.41 300.00     800.00
                            range       se
default                      1.00     0.01
income                  662094.00  3596.04
savings                2911863.00 14006.23
debt                   5968620.00 31046.94
```

r_savings_income	16.11	0.13
r_debt_income	37.00	0.18
r_clothing_income	0.25	0.00
r_education_income	0.53	0.00
r_entertainment_income	1.40	0.00
r_fines_income	0.03	0.00
r_gambling_income	0.21	0.00
r_groceries_income	0.56	0.00
r_health_income	0.40	0.00
r_housing_income	0.43	0.00
r_tax_income	0.09	0.00
r_travel_income	1.40	0.01
r_utilities_income	0.16	0.00
r_expenditure_income	1.33	0.01
cat_gambling	-Inf	NA
cat_debt	1.00	0.01
cat_credit_card	1.00	0.01
cat_mortgage	1.00	0.01
cat_savings_account	1.00	0.00
cat_dependents	1.00	0.01
credit_score	500.00	2.01

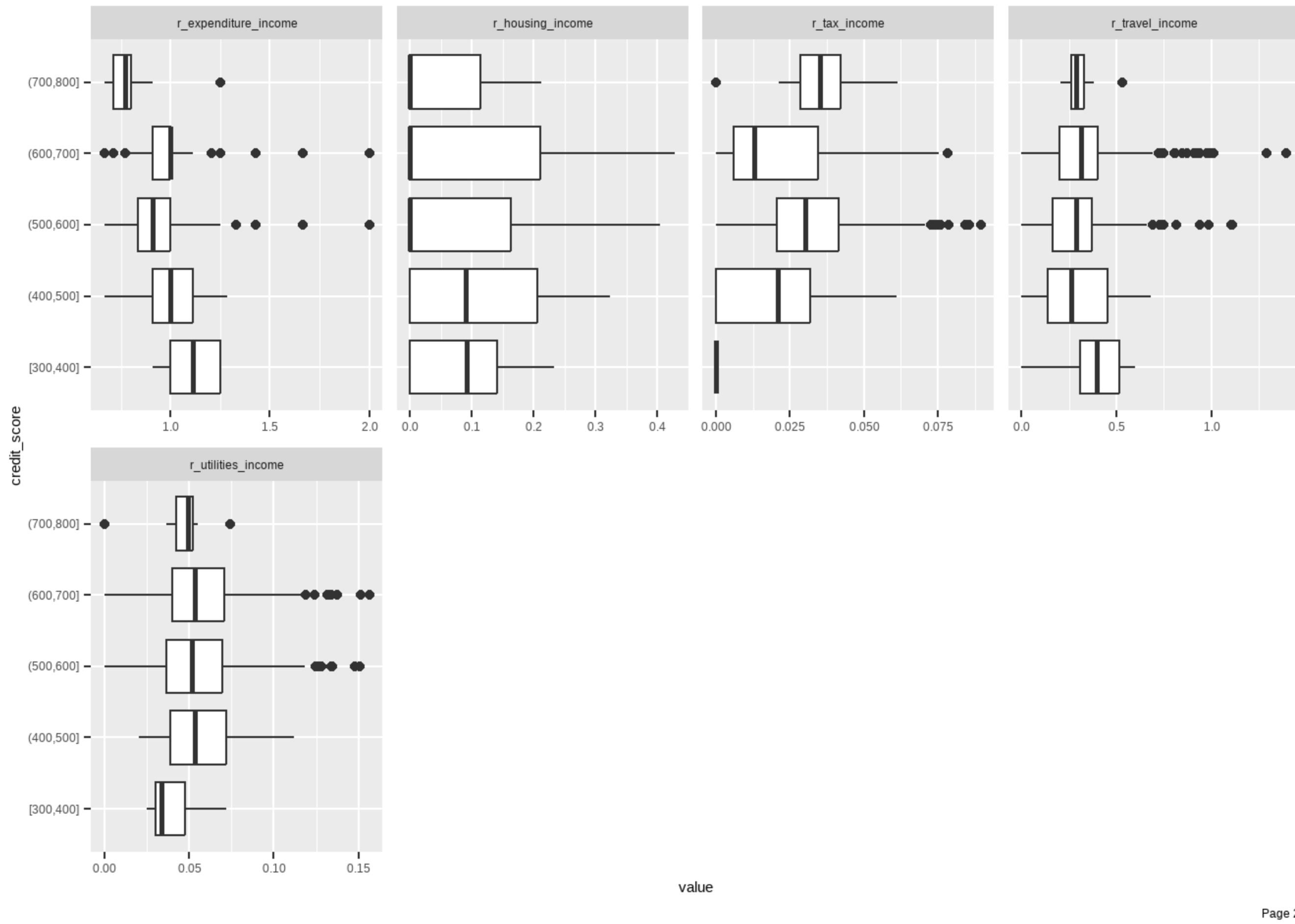
Check summary statistics and variable distributions




credit_score
r_expenditure_income
Frequency
150
100
50
0
300
200
100
0
300
400
500
600
700
800
1.0
1.5
2.0
value



debt
income
r_clothing_income
r_debt_income
r_education_income
r_entertainment_income
r_fines_income
r_gambling_income
r_groceries_income
r_health_income
r_savings_income
savings
credit_score
(700,800]
(600,700]
(500,600]
(400,500]
[300,400]
value



DATA PREPARATION

Address outliers and missing values

It is important to keep outliers in the data and there are no missing values.

Partition the dataset

MODEL DEVELOPMENT

Model 1: Simple Regression

```
Call:
lm(formula = credit_score ~ ., data = trainSet)

Residuals:
     Min       1Q   Median       3Q      Max
-108.882  -16.380    0.624   17.644  153.319

Coefficients:
```

```
                          Estimate     Std. Error t value
(Intercept)            637.591641429  22.135287294  28.804
default                -13.533874544   2.659430558  -5.089
income                   0.000090223   0.000027663   3.261
savings                  0.000003263   0.000005729   0.569
debt                    -0.000006061   0.000002845  -2.130
r_savings_income         0.026099725   0.625827439   0.042
r_debt_income           -8.403108729   0.336106953 -25.001
r_clothing_income      -68.511773338  38.382404533  -1.785
r_education_income      11.782336423  22.877684001   0.515
r_entertainment_income -19.966890222  18.830116268  -1.060
r_fines_income        -145.884709180 851.040747768  -0.171
r_gambling_income        5.053370408  60.828607580   0.083
r_groceries_income     -91.457557547  36.558965633  -2.502
r_health_income       -135.219303475  42.331552695  -3.194
r_housing_income        12.560077428  17.810956687   0.705
r_tax_income           -33.339309310 110.738570347  -0.301
r_travel_income        -14.413590491  14.740066992  -0.978
r_utilities_income     436.889762075 113.571982881   3.847
r_expenditure_income    -3.122050687  15.680209013  -0.199
cat_gamblingLow         20.082183854   5.330745424   3.767
cat_gamblingNo          25.020193283   4.689647580   5.335
cat_debt               -25.945390578   8.820089760  -2.942
cat_credit_card         -1.747768903   3.657666118  -0.478
cat_mortgage            -1.532146559   3.770336042  -0.406
cat_savings_account     13.748851861  17.599311095   0.781
cat_dependents           3.175994075   6.486028808   0.490
                                   Pr(>|t|)
(Intercept)           < 0.0000000000000002 ***
default                        0.000000467 ***
income                            0.001164 **
savings                           0.569208
debt                              0.033508 *
r_savings_income                  0.966747
r_debt_income         < 0.0000000000000002 ***
r_clothing_income                 0.074714 .
r_education_income                0.606712
r_entertainment_income            0.289356
r_fines_income                    0.863946
r_gambling_income                 0.933816
r_groceries_income                0.012598 *
r_health_income                   0.001467 **
r_housing_income                  0.480937
r_tax_income                      0.763459
r_travel_income                   0.328499
r_utilities_income                0.000131 ***
r_expenditure_income              0.842239
cat_gamblingLow                   0.000180 ***
cat_gamblingNo                 0.000000130 ***
cat_debt                          0.003377 **
cat_credit_card                   0.632921
```

```
cat_mortgage                           0.684601
cat_savings_account                    0.434950
cat_dependents                         0.624529
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.58 on 674 degrees of freedom
Multiple R-squared:  0.7921,    Adjusted R-squared:  0.7844
F-statistic: 102.7 on 25 and 674 DF,  p-value: < 0.00000000000000022
```

The simple regression model performs well with an adjusted r-squared value of 0.7844. It has several statistically significant predictors at the .001 level. A key predictor to note is the r_debt_income.

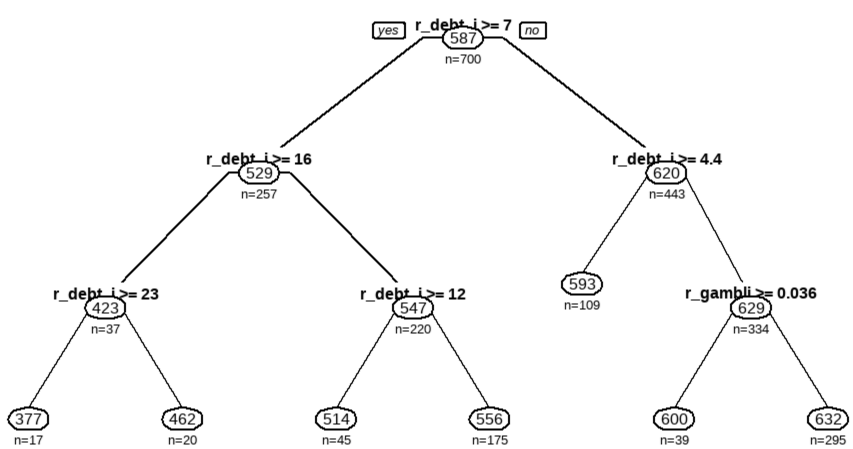
Model 2: Model with Interactions

Using a step wise regression, automatic interactions were discovered and used in the model. This took our model's adjusted r2 from 0.7844 to 0.8521. This now means that more variability can be seen in the model.

Model 3: Regression Tree

Generate the Default Tree

```
DEFAULT TREE
```



The default tree splits 6 times and has 7 nodes. The tree splits based on the ratio of debt to income six times and the the ratio of gambling to income once.

Find the Optimal Tree

```
FULL TREE - Identify the complexity parameter (cp) associated with the smallest cross-validated
prediction error

            CP nsplit rel error    xerror       xstd
1  0.470277520      0 1.0000000 1.0035682 0.07732510
2  0.171842728      1 0.5297225 0.5577879 0.04622593
3  0.036866248      2 0.3578798 0.4023914 0.02766722
4  0.023845233      3 0.3210135 0.3474005 0.02527157
5  0.021957375      4 0.2971683 0.3294517 0.02420626
6  0.013000455      5 0.2752109 0.3141976 0.02374767
7  0.009331006      6 0.2622104 0.3250668 0.02784049
8  0.007398038      7 0.2528794 0.3326371 0.03206760
9  0.006707216      8 0.2454814 0.3348879 0.03203547
10 0.006244437      9 0.2387742 0.3387497 0.03209181
```

The cross-validation error reaches its lowest point at 0.3141976, or 5 splits. A tree with 5 splits is likely optimal, as it achieves the lowest xerror with minimal complexity.

Optimal cp

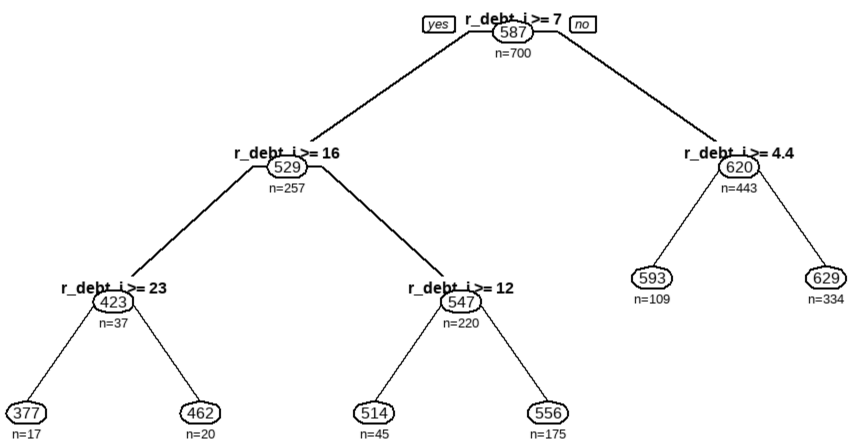
```
[1] 0.013001
```

```
[1] 0.021958
```

We were able to calculate the optimal cp for a tree using the min error method and the best pruned method. In our case, we ended up using the min error cp after further investigation.

Create the Optimal Tree

```
OPTIMAL TREE
```



The optimal tree splits solely on the ratio of debt to income. This tree is organized and does its job well.

MODEL EVALUATION

Model 1: Simple Regression

```
[1] 29.30
```

```
[1] 3.99%
```

After looking at the statistics of the simple regression model, I thought it would be worth it to check the RMSE and the MAPE of it as it already had excellent scores. This gave us a decent RMSE of 29.30 and an excellent MAPE of 3.99%.

Model 2: Regression with Interaction

```
[1] 31.94
```

```
[1] 4.20%
```

The stepwise regression model improved the adjusted r2 significantly. However, after evaluating the RMSE and the MAPE, it is the inferior model compared to the simple regression.

Model 3: Regression Tree – RMSE & MAPE

```
[1] 33.60
```

```
[1] 4.58%
```

The regression tree turned out well, but still not as good as the simple regression. Both the RMSE and the MAPE were close to the other two models, but they were still worse.

DEPLOYMENT

Going back to the questions from the beginning, we can use model 1 to answer each question.
How can financial behaviors such as income, savings, and debt predict an individual's credit score more accurately? Income and debt were both statistically significant, but they were not the most important variables in predicting the credit score. Interestingly, the ratio between the two of them is actually the biggest predictor. Which factors have the greatest impact on predicting credit scores, and how can we leverage these insights for better decision-making? Just like was previously mentioned, the ratio between debt and income has the greatest impact on predicting credit scores. This can be seen in the first model (which is the one we will implement), and also in the other two models. After completing the three, we found that the first model, doing the regression model without interactions, gave us the smallest RMSE and MAPE on the validation set of data. The MAPE is very encouraging at 3.99%, however, the RMSE does raise some concerns at 29.30. Credit score should not vary that heavily, so that is something that will need to be looked at closer before implementing the model. We are on the right track for better predicting a customers credit score, and once more fine tuned, this will lead to better insights into key factors that drive a person to default in the classification model.

REFERENCES

R and Packages

```
R version 4.3.1 (2023-06-16 ucrt)
```

```
R Packages Used:

 [1] "readxl"       "janitor"      "dplyr"        "summarytools" "corrplot"
 [6] "dlookr"       "knitr"        "formattable"  "DataExplorer" "MASS"
[11] "forecast"     "rpart.plot"   "rpart"        "caret"        "lattice"
[16] "ggplot2"
```

Other References

Jaggia, S., Kelly, A., Lertwachara, K., & Chen, L. (2023). *Business analytics: Communicating with numbers* (2nd Ed.). McGraw-Hill.