5596 -Data Analytics

R for Marketing

Group 5: Adrian Lehrner, Pascal Jakub Schreiber, Henriett Kapui, Marta Wiśniewska, Sebastian Frey

Table of content

Motivation and research question

What is the estimated income of customers who did not state it?

- Income prediction for marketing purposes
- Importance of income predictors such as: Education, Marital Status, Age

Our data

Summary

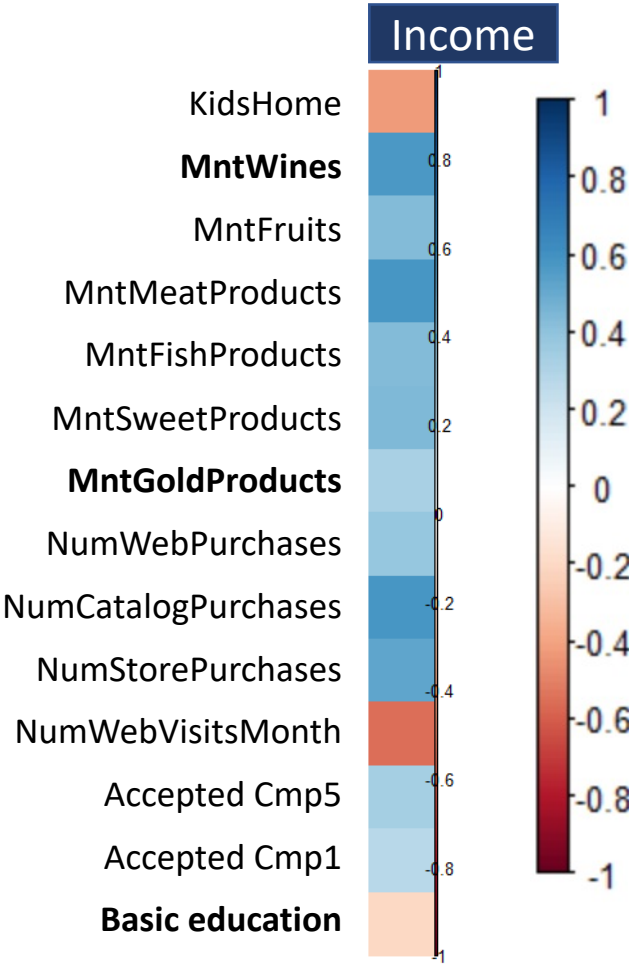
Our dataset consists of customer's data collected since their enrollment with the company.

- Source:
  - https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign
- Size: 220 KB
- Format: CSV
- Variables: 2240 Rows x 29 Columns
- Pre-processing:
  - Removing attributes not related to customers
  - Creating new columns
  - Splitting to train and test

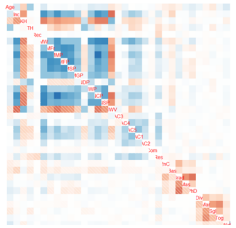
PCA, Correlation matrix

Only variables with a Correlation > 0.2 are considered

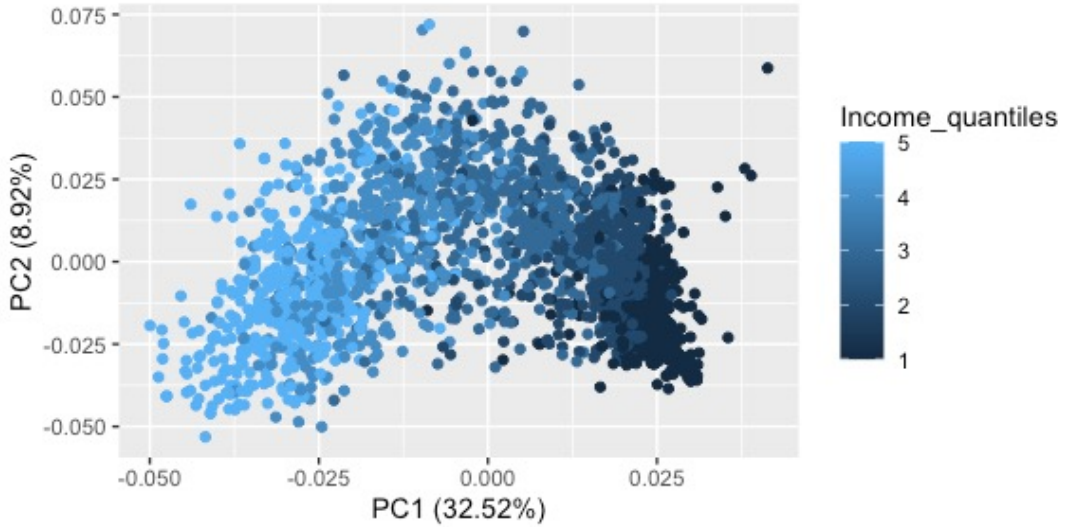
Interpretation might not be that straight forward/intuitive for some of the variables but it does appear reasonable that **Gold purchases** and **Wine purchases** correlate positively with Income while **Basic education** has a negative effect



PCAs vs. 5 Income quantiles



Lets simplify



After **reducing the complexity** of the data through PCA a clear **pattern** of different **subgroups** regarding income is **visible**

Comparing data via a classification approach

KNN

- Overall better performance of KNN
- If income is split into a binary variable accuracy of 91%
- If income is split into 5 quantiles reduction of accuracy, but still accuracy of 73%

Naive Bayes

- Naive Bayes approach is worse than KNN in every metric except precision, here they are approximately equal
- KNN is preferable over Naive Bayes

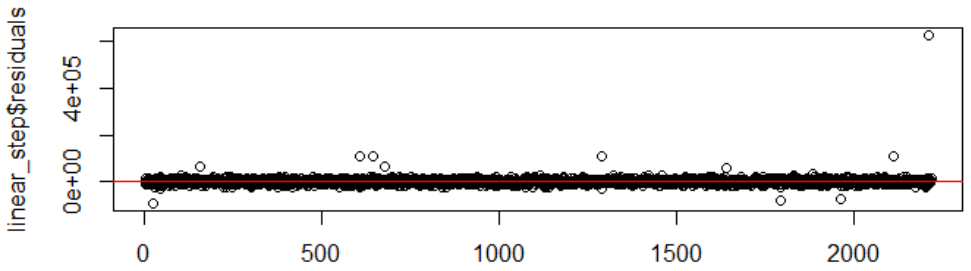
Linear Regression

Full model with all variables

```
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         45982.311   2707.796  16.981  < 2e-16 ***
Age                    28.123     34.128   0.824 0.410006
Kidhome              2080.657    923.041   2.254 0.024286 *
Teenhome             5014.257    838.256   5.982 2.57e-09 ***
Recency               -20.170     12.819  -1.573 0.115775
MntWines               11.996      1.982   6.051 1.68e-09 ***
MntFruits              22.941     12.659   1.812 0.070082 .
MntMeatProducts        18.709      2.798   6.686 2.90e-11 ***
MntFishProducts         7.131      9.642   0.740 0.459616
MntSweetProducts       24.891     12.214   2.038 0.041684 *
MntGoldProds           -4.515      8.653  -0.522 0.601844
NumDealsPurchases    -372.107    242.108  -1.537 0.124451
NumWebPurchases      1114.077    183.984   6.055 1.65e-09 ***
NumCatalogPurchases   831.851    216.444   3.843 0.000125 ***
NumStorePurchases     405.044    173.134   2.339 0.019400 *
NumWebVisitsMonth   -3022.807    219.758 -13.755  < 2e-16 ***
AcceptedCmp31       -1790.836   1483.894  -1.207 0.227621
AcceptedCmp41        3272.810   1617.932   2.023 0.043212 *
AcceptedCmp51        3680.767   1791.992   2.054 0.040094 *
AcceptedCmp11        3096.491   1717.037   1.803 0.071464 .
AcceptedCmp21        1596.658   3363.341   0.475 0.635031
Complain1           -1189.784   3735.422  -0.319 0.750126
Response1            -765.738   1216.375  -0.630 0.529072
Basic1             -10680.090   2642.908  -4.041 5.51e-05 ***
Graduation1          1462.140   1314.422   1.112 0.266096
Master1              1608.467   1533.547   1.049 0.294361
PhD1                 2872.607   1506.302   1.907 0.056644 .
Divorced1             843.598   1372.567   0.615 0.538874
Married1              129.083    982.798   0.131 0.895516
Together1            1106.500   1068.981   1.035 0.300738
Widow1               -481.912   2137.076  -0.226 0.821611
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16940 on 2185 degrees of freedom
Multiple R-squared:  0.5535,    Adjusted R-squared:  0.5473
F-statistic: 90.27 on 30 and 2185 DF,  p-value: < 2.2e-16
```

Stepwise model by AIC

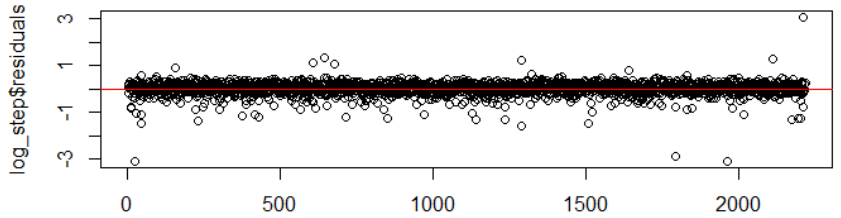
```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         49341.759   1830.826  26.951  < 2e-16 ***
Kidhome              2081.903    908.137   2.292   0.0220 *
Teenhome             5382.190    790.499   6.809 1.27e-11 ***
Recency               -18.564     12.460  -1.490   0.1364
MntWines               12.927      1.923   6.723 2.26e-11 ***
MntFruits              21.735     12.126   1.793   0.0732 .
MntMeatProducts        18.754      2.739   6.848 9.69e-12 ***
MntSweetProducts       23.447     11.793   1.988   0.0469 *
NumDealsPurchases    -427.225    239.885  -1.781   0.0751 .
NumWebPurchases      1103.313    180.072   6.127 1.06e-09 ***
NumCatalogPurchases   840.453    213.374   3.939 8.44e-05 ***
NumStorePurchases     397.645    170.616   2.331   0.0199 *
NumWebVisitsMonth   -3079.315    215.936 -14.260  < 2e-16 ***
AcceptedCmp31       -2157.728   1434.846  -1.504   0.1328
AcceptedCmp41        3200.444   1568.005   2.041   0.0414 *
AcceptedCmp51        3246.867   1744.291   1.861   0.0628 .
AcceptedCmp11        2939.936   1683.101   1.747   0.0808 .
Basic1             -12238.577   2392.764  -5.115 3.41e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16920 on 2198 degrees of freedom
Multiple R-squared:  0.5518,    Adjusted R-squared:  0.5483
F-statistic: 159.2 on 17 and 2198 DF,  p-value: < 2.2e-16
```



Linear Regression

Log-Linear Stepwise model by AIC

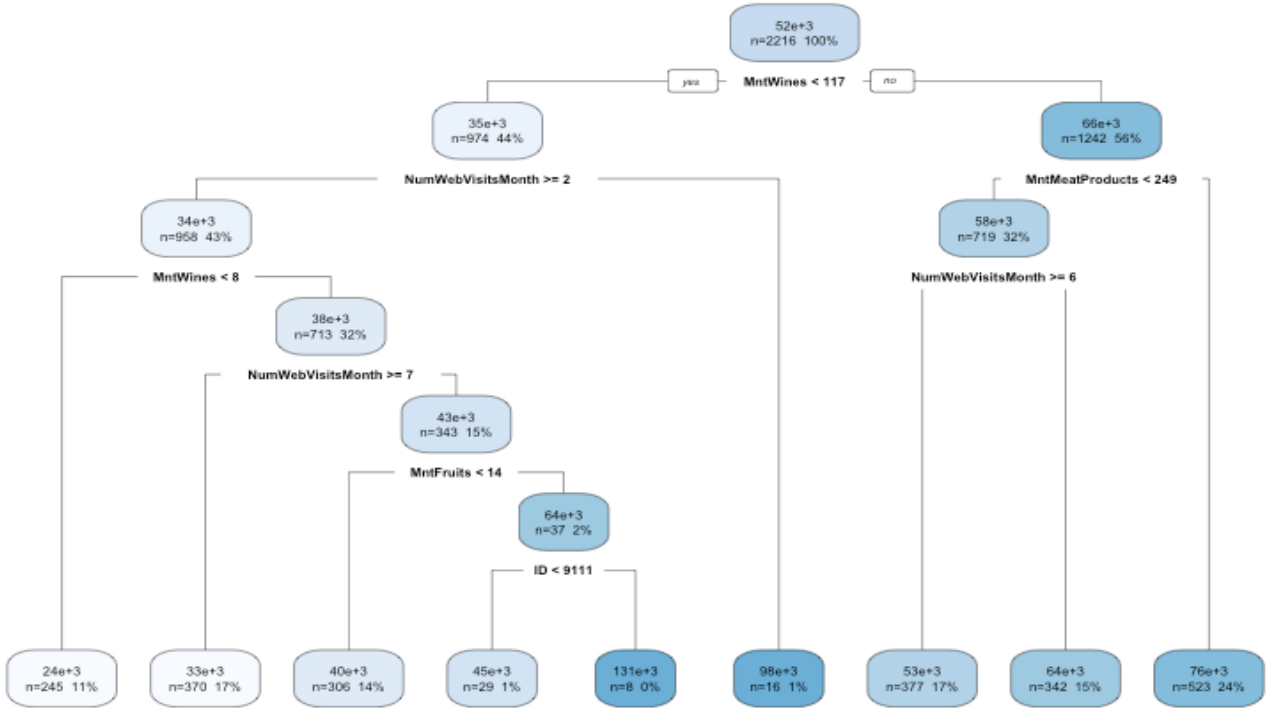
```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.054e+01  4.286e-02 245.864  < 2e-16 ***
Age                  1.354e-03  5.640e-04   2.400  0.01648 *
Kidhome              8.846e-02  1.539e-02   5.749 1.02e-08 ***
Teenhome             1.797e-01  1.407e-02  12.771  < 2e-16 ***
MntWines             3.120e-04  3.128e-05   9.973  < 2e-16 ***
MntFruits            3.984e-04  2.118e-04   1.881  0.06009 .
MntMeatProducts      2.475e-04  4.649e-05   5.324 1.12e-07 ***
MntFishProducts      3.512e-04  1.594e-04   2.204  0.02763 *
MntSweetProducts     4.271e-04  2.044e-04   2.089  0.03682 *
NumDealsPurchases   -2.247e-02  4.025e-03  -5.582 2.67e-08 ***
NumWebPurchases      2.944e-02  3.041e-03   9.681  < 2e-16 ***
NumCatalogPurchases  6.351e-03  3.567e-03   1.781  0.07512 .
NumStorePurchases    2.148e-02  2.884e-03   7.450 1.34e-13 ***
NumWebVisitsMonth   -7.563e-02  3.653e-03 -20.703  < 2e-16 ***
AcceptedCmp41        7.550e-02  2.577e-02   2.929  0.00344 **
Response1            3.033e-02  1.848e-02   1.641  0.10096
Basic1              -3.769e-01  4.441e-02  -8.486  < 2e-16 ***
Graduation1          3.780e-02  2.211e-02   1.710  0.08749 .
Master1              5.795e-02  2.573e-02   2.252  0.02440 *
PhD1                 7.770e-02  2.511e-02   3.094  0.00200 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2852 on 2196 degrees of freedom
Multiple R-squared:  0.6848,    Adjusted R-squared:  0.6821
F-statistic: 251.1 on 19 and 2196 DF,  p-value: < 2.2e-16
```

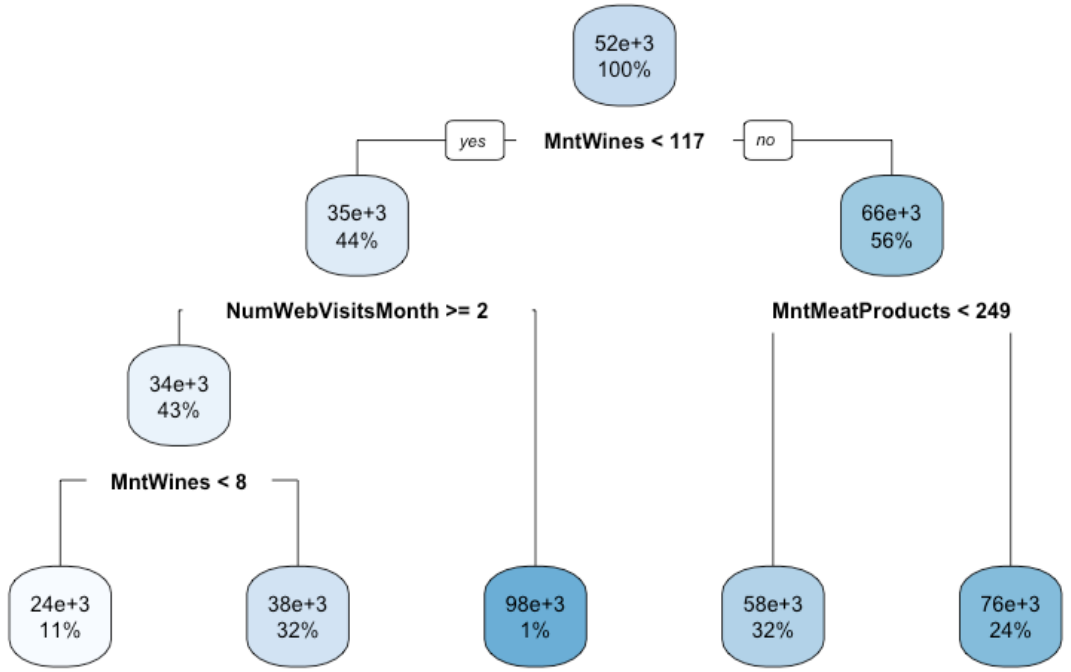


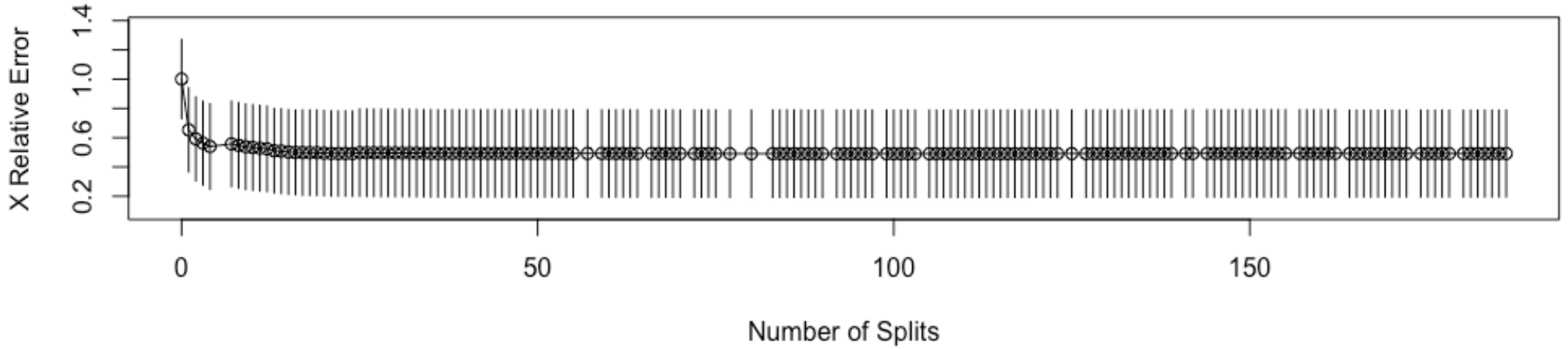
Comparison test/train set

Model	AIC	BIC	Adj R^2	RMSE
Full with all variables	49476,93	49659,44	54,73%	17948.93
Log-Linear stepwise by AIC	751,04	870,81	68,21%	18546.47
Stepwise by AIC	49459,34	49567,71	54,83%	17805.81

Regression trees

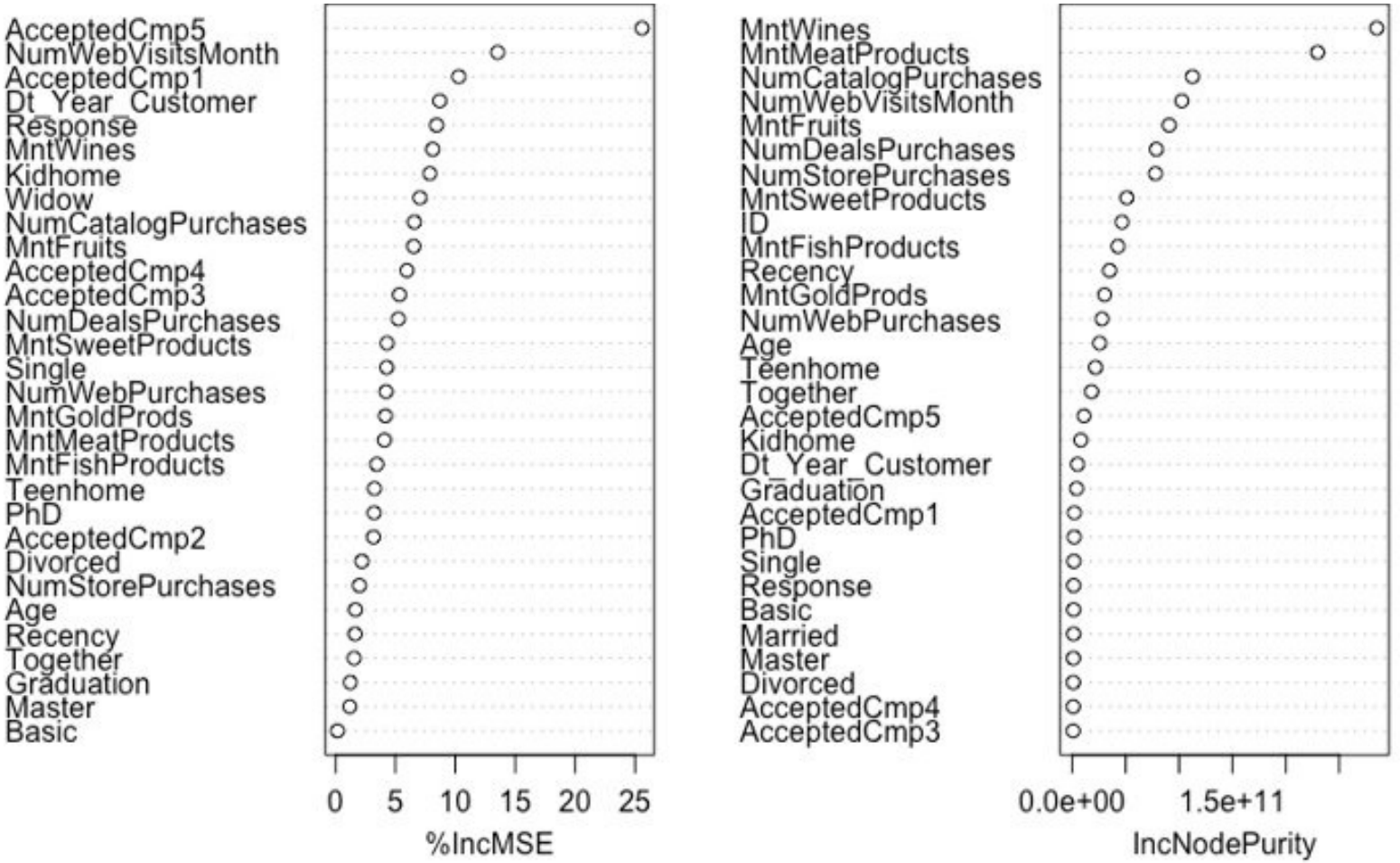






Random forests

```
Call:
 randomForest(formula = Income ~ ., data = marketing_dat_clean,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 11

          Mean of squared residuals: 277696907
                    % Var explained: 56.16
```



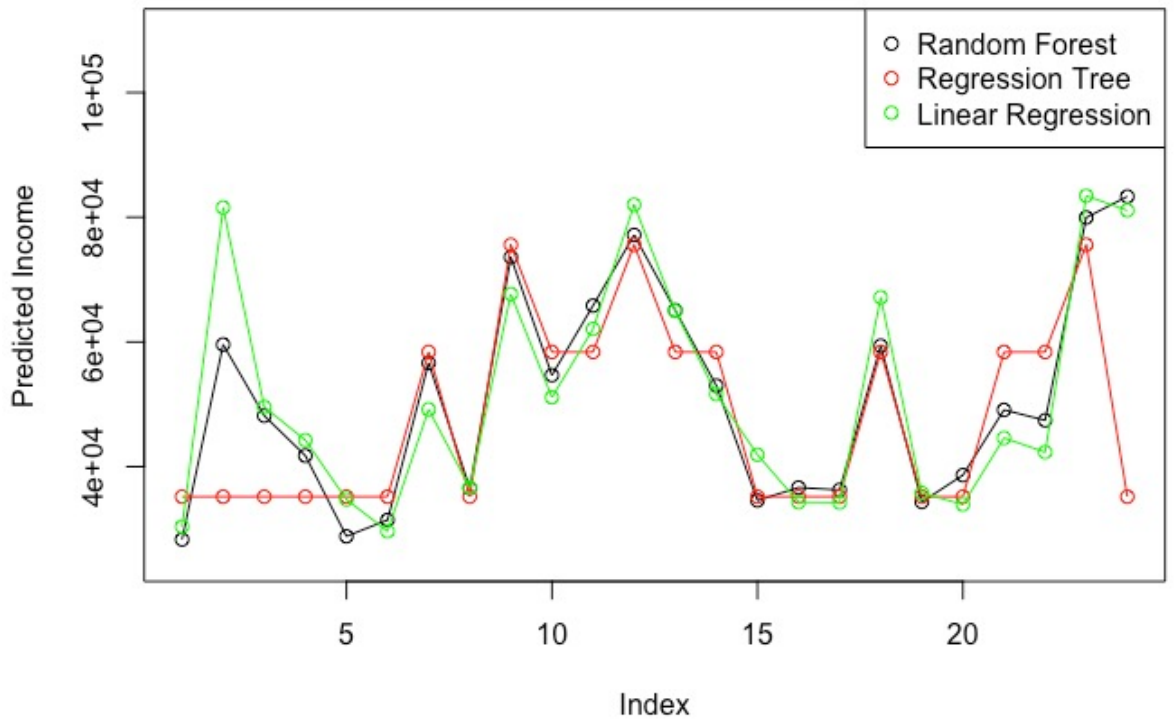
Model comparison

Regression tree, Random Forest and Linear regression

Model	RMSE (Goodness-of-fit)	R-squared (% of Variance)
Random Forest	12657.07	0.7550511
Regression tree	16978.94	0.5320547
Linear regression	13553.14	0.7065903

Winning Model

We chose the Random Forests model for the prediction of the missing income values.



Thank you for your attention!