
# Report Open Analysis

**Puerta de Toledo**

**Master in Big Data Analytics**

**Predictive Modeling**

Ion Bueno Ulacia NIA 100364530

Daniel Gil Santiuste NIA 100364564

Daniel Martín Cruz NIA 100384121

Ion Bueno Ulacia NIA 100364530
Daniel Gil Santiuste NIA 100364564
Daniel Martín Cruz NIA 100384121

## 1. Introduction

The amount of money that anyone is willing to pay for a product is the base for any market. This is not an exception for vehicle market, where understanding the dynamics and patterns of the own market brings huge benefits on decision making and determines the effectiveness of the long-term investment.

Hence, it is important to develop appropriate predictive models to identify the consumer characteristics. This is a crucial task in the current furious market where e-commerce, direct sales and new vehicle models play a very important role. As a consequence of understanding this market and specially the role of pricing, a maximization of the benefits is expected.

This analysis attempts to study the car market purchasing trends for our company. In this spirit, different statistical methodologies are applied in order to quantify the relations between car prices and population features. According to this objectives, the following questions are addressed in the present study:

- Can we predict the price someone is willing to pay effectively?
- Which is the influence of the predictors in the response and how can they be interpreted?

Then the study is set to develop a model to predict the amount that our clients are willing to pay for a new car according to their social features. In this way, trends on gender or age could be found. This information will be used to establish the market targets to meet the most interesting customer profiles.

The knowledge behind the deployed study is on the incredible notes from [1], which have been used as a reference and the guideline for the steps of the present study. From this point of view, all the credit goes for them.

## 2. Statistical analysis

**Dataset**

This study will be develop based on the information collected in [2]. Due to our company data policy and to respect the customers privacy, the personal information concerning them (name and e-mail) are removed before performing any analysis.

Then the dataset contains five numerical variables:

- Age
- Salary
- Card debt
- Net worth
- Purchase amount: target variable. Indicates the price for which the car in particular was sold.

And one categorical variable:

- Gender: already codified as "0" for women and "1" for men.

The size of the dataset is 500 samples. In order to obtain the predictive model and for the study results to be consistent, 20 of these samples are taken apart and reserved for the last step of checking whether the model is capable of effectively predicting the appropriate car price.

### Data visualization

In order to get a first idea of the structure of the data, it is a good practice to obtain a useful visualization of the dataset which is going to be handled. In this line of thought, a plot that could give valuable insights is the scatter plot matrix where possible relations between variables could be observed:

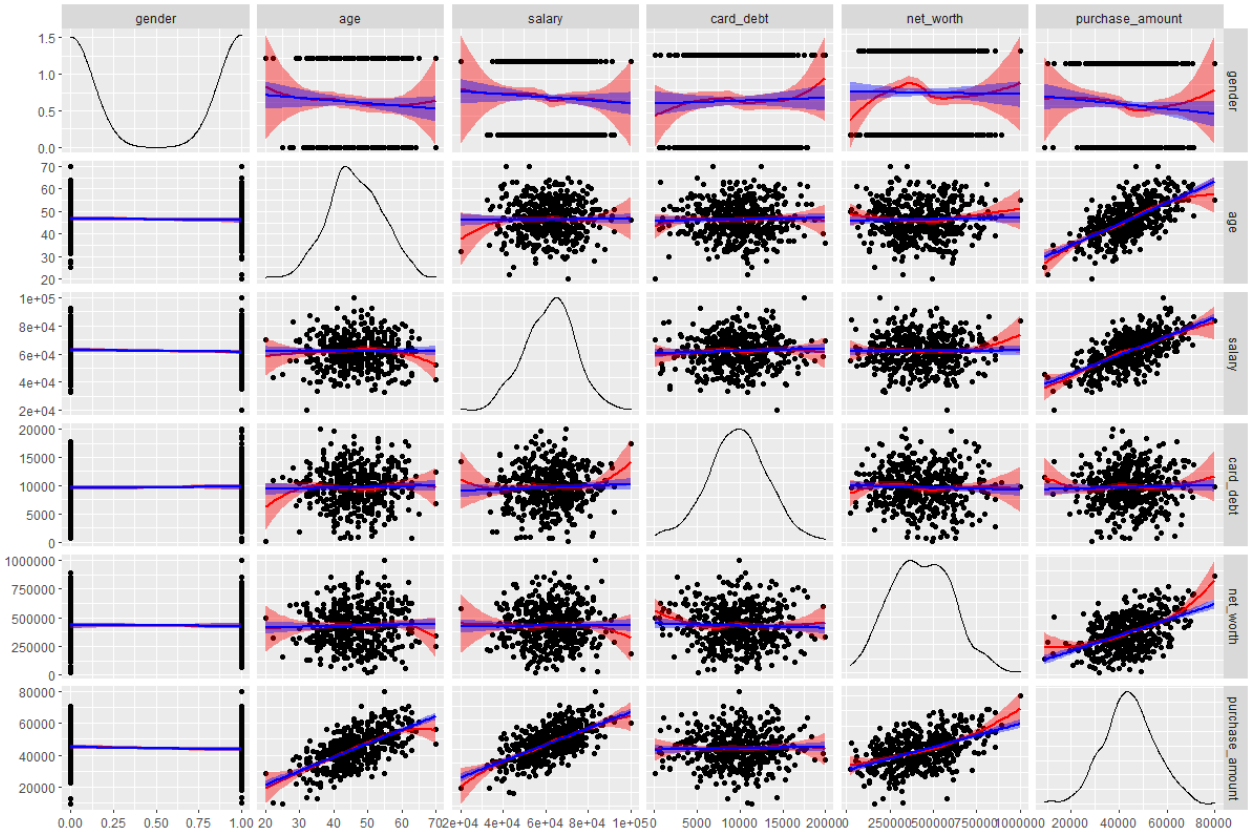



Figure 1: Scatter plot matrix

As it can be seen, no pair of variables shows a clear dependence between them. This fact discards the option of removing any of the predictors at this point of the process. Nevertheless, it is assumed that some predictors are going to be more important than others.

### Simple linear regression models

The representation in figure 1 also shows that the most evident relation between a predictor and the variable to predict (*purchase_amount*) could be the one between *age* or *salary* with this *purchase_amount* variable. These variables seem to be the ones which could give the most trustworthy simple linear regression model. Although a more complex model is willing to be obtained, this could be a good starting point and so, these models are tried (figures 2a and 2b).

```
Call:
lm(formula = purchase_amount ~ age, data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-23214.4  -5689.8    321.5   5517.0  28321.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4159.55    2241.30   1.856   0.0641 .
age           863.97      47.68  18.119   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8343 on 478 degrees of freedom
Multiple R-squared:  0.4072,    Adjusted R-squared:  0.4059
F-statistic: 328.3 on 1 and 478 DF,  p-value: < 2.2e-16
```

(a) Simple linear model for age

```
Call:
lm(formula = purchase_amount ~ salary, data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-25545.2  -5834.2   -681.1   5965.1  23764.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8845.9229  2100.7096   4.211 3.04e-05 ***
salary         0.5699     0.0333  17.117  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8532 on 478 degrees of freedom
Multiple R-squared:   0.38,     Adjusted R-squared:  0.3787
F-statistic:   293 on 1 and 478 DF,  p-value: < 2.2e-16
```

(b) Simple linear model for salary

Figure 2: Simple linear models

The $R^2$ obtained is equal to 0.4072 for the model on age and 0.38 in the case of the salary model. From the results, it is clear that these models do not properly explain the variable of interest in this study. Nevertheless, they can give some clues about the influence of the predictors and their relationship with the variable to predict. This will be seen as the model gets more complicated in the following steps of the process.

**Linear models with more predictors**

Clearly, the next step should be including more predictors in the previously generated models. In this way, the $R^2$ and the prediction accuracy for the purchase amount prediction should increase. Hence, a new model this time including all the available predictors is generated and can be found in figure 3a.

```
> modAll <- lm(purchase_amount ~ ., data = cars)
> summary(modAll)

Call:
lm(formula = purchase_amount ~ ., data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-465.11 -204.89   19.71  200.31  435.52

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -4.211e+04  9.592e+01 -439.036   <2e-16 ***
gender       3.024e+01  2.213e+01    1.367   0.1724
age          8.399e+02  1.385e+00  606.636   <2e-16 ***
salary       5.626e-01  9.457e-04  594.846   <2e-16 ***
card_debt    5.214e-03  3.157e-03    1.652   0.0993 .
net_worth    2.893e-02  6.335e-05  456.688   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 241.7 on 474 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9995
F-statistic: 1.92e+05 on 5 and 474 DF,  p-value: < 2.2e-16
```

(a) Linear model using all predictors

```
> modAllSig <- lm(purchase_amount ~ . -card_debt -gender, data = cars)
> summary(modAllSig)

Call:
lm(formula = purchase_amount ~ . - card_debt - gender, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-438.75 -215.29   21.91  201.21  421.68

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.204e+04  9.085e+01  -462.8   <2e-16 ***
age          8.399e+02  1.386e+00   605.9   <2e-16 ***
salary       5.626e-01  9.461e-04   594.6   <2e-16 ***
net_worth    2.893e-02  6.347e-05   455.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 242.4 on 476 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9995
F-statistic: 3.182e+05 on 3 and 476 DF,  p-value: < 2.2e-16
```

(b) Linear model using only significant predictors

Figure 3: Linear models with all variables

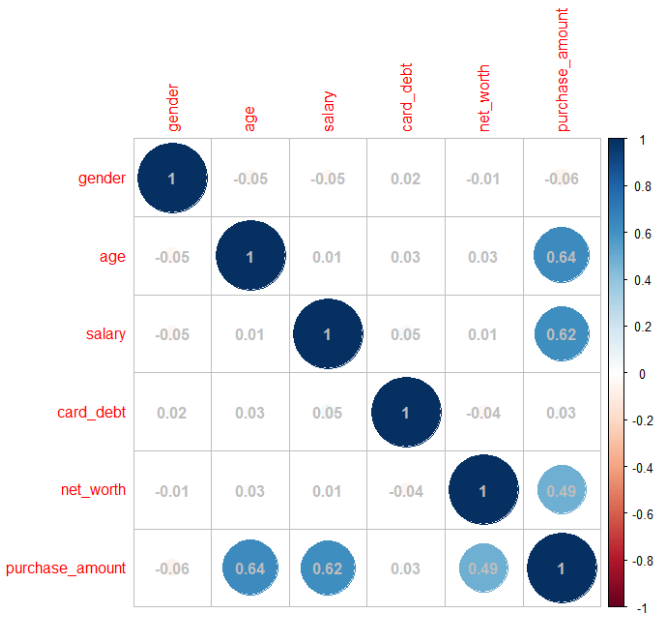
As it can be seen, now $R^2$ is clearly higher with respect to the simple linear regression models. Then, one of the first conclusions would be that this combination of predictors leads to important improvements in the prediction task which is being performed.

A deeper analysis for the model should be carried out. From the summary of the model, predictors *gender* and *car_debt* can be declared as non-significant. This tells that there exists an excess of predictors which unnecessarily complicate the model. It was tried to remove just one of them and, after checking the resulting

model, in any case the other predictor was still non-significant. Then both variables, *gender* and *car_debt* were removed resulting in the model from figure 3b.

The reason for removing *gender* and *car_debt* can be statistically studied. It could be due to an existing correlation with other variables. In any case, it was not appreciated in the scatterplot of the variables. Then the correlation matrix is plotted as figure 4a and it can be seen that this is not the reason we are looking for.

Nevertheless, as multicollinearity could exist and it is not enough to inspect just pairwise correlations in order to get rid of it, the VIF of each coefficient of the model with all the predictors is calculated. Their values, very close to 1 (figure 4b), tell that multicollinearity is not the problem neither.




(a) Correlation plot of the variables

```
> car::vif(modAll)
    gender        age     salary card_debt net_worth
1.005354   1.004234   1.004956   1.005683   1.002748
```

(b) VIF calculation

Figure 4: Variables relation study

After this approach, it could be thought that removing these predictors from the final linear model is just a consequence of the own problem nature. The fact that the coefficient corresponding to gender is non-significant and so removed from the final model gives an interesting conclusion about the car purchasing market. This is that gender has no influence in cars acquisition. In this line of analysis, the corresponding simple linear regression model can be studied to check the validity of such a conclusion. This is done because *gender* was considered a relevant variable before starting the study and the consequences of not including it in the final model should be studied carefully.

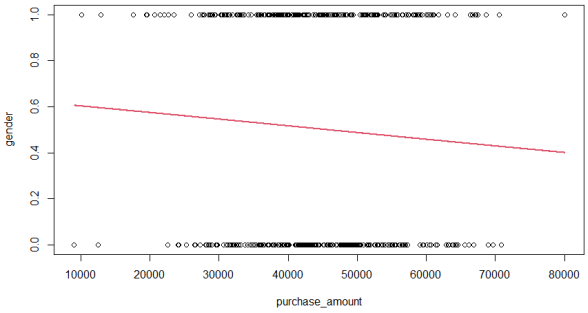
```
Call:
lm(formula = purchase_amount ~ gender, data = cars)

Residuals:
   Min     1Q Median     3Q    Max
-35866  -6494   -445   7097  36496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44866.0      701.0  64.004   <2e-16 ***
gender       -1361.7      987.2  -1.379    0.168
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10810 on 478 degrees of freedom
Multiple R-squared:  0.003964,  Adjusted R-squared:  0.00188
F-statistic: 1.902 on 1 and 478 DF,  p-value: 0.1685
```

(a) Simple linear model for gender




(b) Logistic regression with gender

Figure 5: Gender analysis

The meaning of this model, shown in figure 5a is that men pay 1.361,7 dollars less when purchasing a car. Nevertheless, even in this case where it would be the only predictor, it is considered as non-significant.

However, a linear model is not the best approach for this variable, since it can only take values 0 or 1. For this reason, a logistic regression is tried, in order to see if it the *purchase_amount* is related with the *gender*. Nevertheless, as it can be seen in figure 5b, the fitting is very poor, concluding that variables are not correlated.

This analysis says that variable *gender* should be clearly not taken into account in the model.

**ANOVA decomposition**

Once this analysis has been done, the models with more predictors are recovered. The shown linear dependence can be more deeply analyzed by the ANOVA decomposition of the models in figures 6a and 6b.

```
> simpleAnova(modAll)
Analysis of Variance Table

Response: purchase_amount
            Df     Sum Sq    Mean Sq F value    Pr(>F)
Predictors   5 5.6097e+10 1.1219e+10  191991 < 2.2e-16 ***
Residuals  474 2.7699e+07 5.8437e+04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Linear model using all predictors

```
> simpleAnova(modAllSig)
Analysis of Variance Table

Response: purchase_amount
            Df     Sum Sq    Mean Sq F value    Pr(>F)
Predictors   3 5.6097e+10 1.8699e+10  318170 < 2.2e-16 ***
Residuals  476 2.7974e+07 5.8770e+04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(b) Linear model using only significant predictors

Figure 6: ANOVA decomposition

As the tables for the ANOVA decomposition show, the hypothesis of no linear dependence of *purchase_amount* on the given predictors is rejected for both models. Then at least one of the predictors' coefficients will be different from 0. Starting from the model which includes all the variables from the dataset as predictors (figure 6a), the idea of the performed test still holds when removing the non-significant variables (figure 6b).

In fact, this model where non-significant predictors are removed (presented in figure 3a) should be kept as the final result of this selection procedure. In this way, a model where there should be no redundant information has been obtained.

This manual process that has been followed could be done in a more proper way. The best subset of predictors

could be obtained by applying stepwise model selection and choosing the best one in terms of BIC which turns to be a metric which penalizes more the complex model than AIC when performing model comparison. Anyway, any of them is a more trustworthy metric than $R^2$ for selecting model. Applying this technique leads to exactly the same model which was obtained by removing the non-significant predictors by hand.

**Non-linear extensions**

Looking for an improvement of the model, it is also tried to extend the study to look for non-linear relations of the predictors with the variable of interest. In this search for the best model, first-order interactions among the predictors are included. From this process, it can be stated that no improvement can be made by including any terms interaction as it results in the same best model as obtained until this point of the study.

Mention a specific study has been performed for the interactions with *gender*, in order to see if two populations could be distinguished by this variable. Remark that the goal was different than before with the logistic regression, where the idea was looking into the direct explanation of *purchase_amount* by gender. However, the interactions with *gender* were removed performing stepwise model selection, what makes sense with figure 7, since the points of both genders are very well mixed.

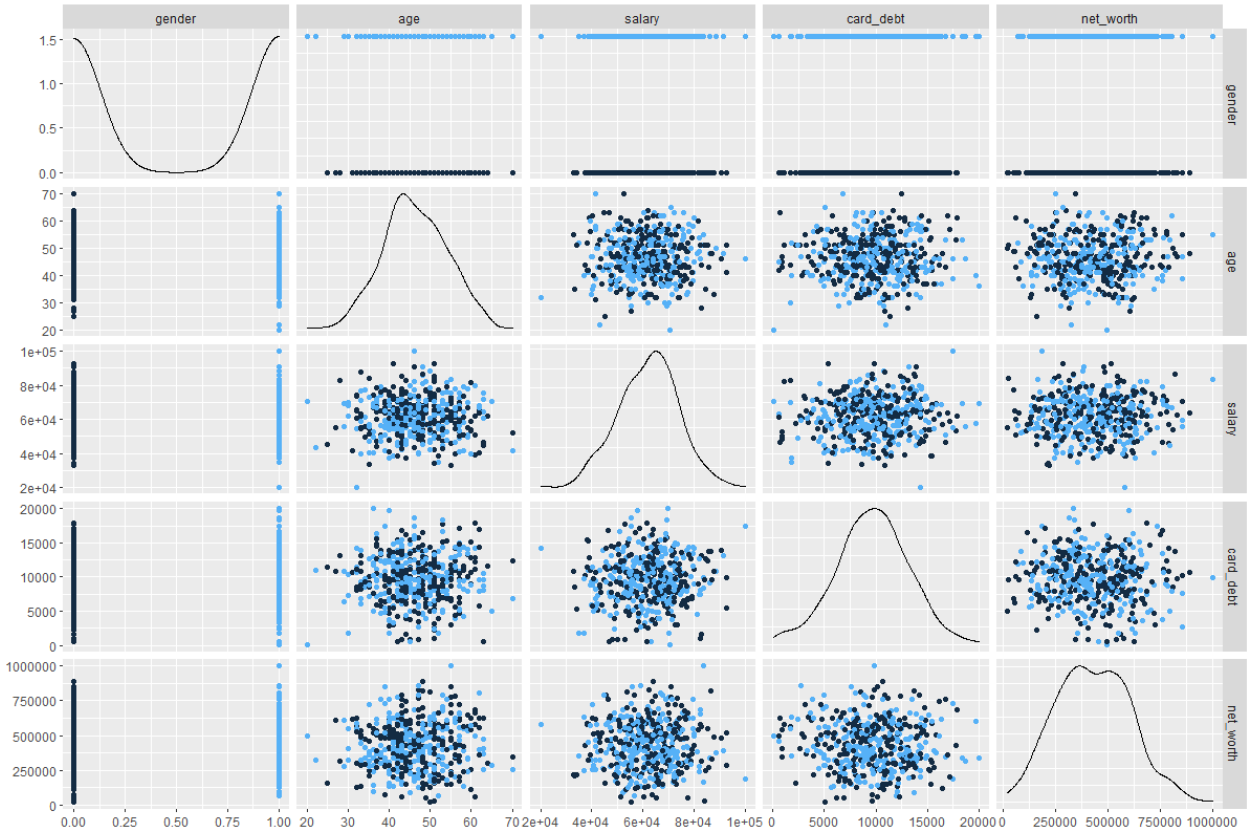


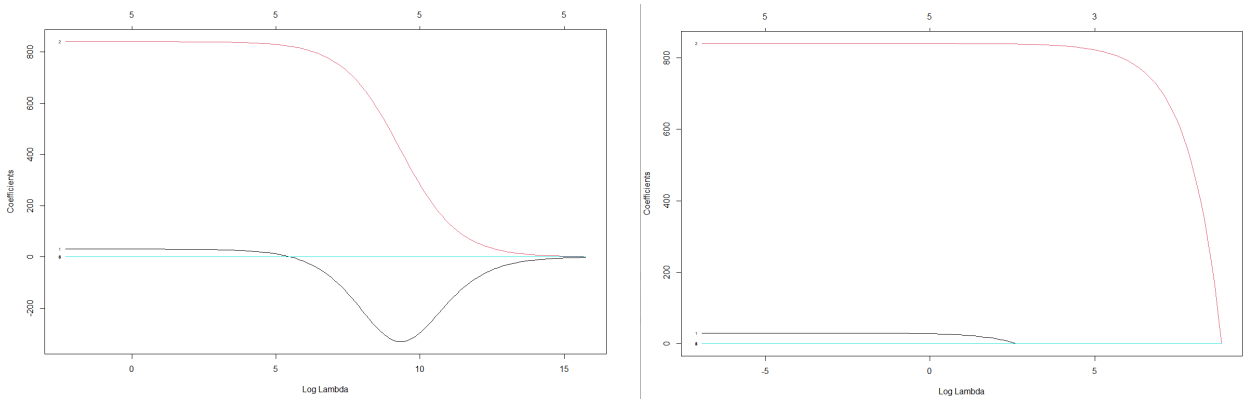

Figure 7: Scatter plot matrix differentiating between genres

### Ridge and Lasso regression

Alternatively, Ridge and Lasso regression models are also included in this model selection process in order to try a wide range of different approaches.

The ultimate goal of these two regression is to perform variable selection on our original data and check if we obtain a similar result as we did before, concluding that *gender* and *car_debt* are not valuable variables in order to predict the price of the cars sold. We know that the proper way of doing this variable selection is by applying Lasso and then using the `stepAIC` function but given the reduced amount of predictors present in this dataset, we decided proceeding in this way. In addition, the results obtained using both approaches independently match.

The process of parameter tuning is further explained in the R script, in which we obtain the following coefficient for our variables as the value of lambda increases:




(a) Coefficients using Ridge regression

(b) Coefficients using Lasso regression

Figure 8: Variables coefficient vs. log(Lambda) value

Here it can be seen that both models coincide in the conclusion that has been already mentioned, that only three variables (*salary*, *age* and *net_worth*) are sufficient to explain the target variable *purchase_amount*.

The are some characteristics to be highlighted according the figure 8. In the figure 8a, we can see that the "useless" variables approach to 0 as lambda goes up, not reaching this value in any moment. On the other hand, in the case of Lasso regression depicted in figure 8b, these "useless" predictor reach 0 when the value of log(Lambda) is between 0 and 5, discarding those two variables, as it can be seen in the upper part of the figure. This property of discarding variables makes Lasso as great method for performing variable selection.

### Prediction

Once the model selection process has ended up, it is time to test the final model in a prediction scenario, as its prediction capacity was one of the questions presented as the study objectives. For this purpose, the samples which were taken apart just at the beginning of the study (in order to have no influence in the model composition) are used in this prediction task and tried to be predicted. These results are shown in figure 9.

```
> preds_vs_true
         fit      lwr      upr   y_true
57  47575.93 47096.03 48055.84 47380.91
87  44774.78 44297.81 45251.76 45167.33
96  40274.91 39797.90 40751.92 40004.87
159 39326.65 38848.32 39804.98 39433.41
189 61124.60 60646.45 61602.74 60865.76
203 41031.35 40553.51 41509.18 40660.38
210 41893.89 41416.25 42371.53 42209.29
226 27834.26 27355.02 28313.50 27625.44
227 46401.12 45923.30 46878.93 46389.50
238 49639.36 49162.23 50116.50 49399.97
250 46059.02 45580.38 46537.66 46135.27
257 48421.74 47943.42 48900.07 48349.16
269 41225.84 40747.70 41703.99 41320.07
318 59994.12 59514.59 60473.65 59758.73
335 52502.15 52024.98 52979.32 52240.73
396 30519.34 30041.21 30997.46 30757.66
421 61755.62 61276.45 62234.80 62028.71
425 31131.62 30653.64 31609.59 31408.63
446 52497.98 52019.45 52976.50 52150.42
460 35344.57 34866.68 35822.46 35457.15
```

Figure 9: Prediction results

The first impression on these results is the similarity between columns *fit* and *y_true*, which refer to the predicted and real values for each of the samples, respectively. Going deeper into this analysis, it can be seen that, in fact, the predictions (*y_true*) are always inside the confidence intervals (with $\alpha$ level equal to 0.05) determined by the lower (*lwr*) and the upper (*upr*) bounds. This behaviour proves an outstanding performance of the developed model.

Another consideration should be made on the intervals length. They are not wide larger than 1.000 dollars in any of the predicted samples. This is, again a really good indicator of the predictions quality as car prices, in mean, are around 44.000 dollars.

## 3. Conclusions

The model `modAllSig` is the one with the lowest BIC, which was the chosen metric for the model selection. This same result was obtained for the Lasso regressor with $\hat{\lambda}_{n-1SE}$, but the first approach by using regular linear models is considered simpler as so it is the one used as final model. The summary of this model can be found in figure 10.

```
> modAllSig <- lm(purchase_amount ~ . -card_debt -gender, data = cars)
> summary(modAllSig)

Call:
lm(formula = purchase_amount ~ . - card_debt - gender, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-438.75 -215.29   21.91  201.21  421.68

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.204e+04  9.085e+01  -462.8   <2e-16 ***
age          8.399e+02  1.386e+00   605.9   <2e-16 ***
salary       5.626e-01  9.461e-04   594.6   <2e-16 ***
net_worth    2.893e-02  6.347e-05   455.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 242.4 on 476 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9995
F-statistic: 3.182e+05 on 3 and 476 DF,  p-value: < 2.2e-16
```

Figure 10: Final model

As it can be seen, this model explains the 99.95% of the variability in a non-redundant way and without non-significant coefficients. In this way, the first conclusion would be that a formula that effectively explains and predicts the car purchasing price has been obtained. This is supported by the prediction test whose results are shown in figure 9. Also the uncertainty of the predictions was proved to be good enough for the values we are working with in this study. And then the first one of the proposed questions is answered.

The second question stands for the model coefficients interpretation:

- Older people spends higher money quantities when purchasing a car.
- These purchaising quantities are also higher as people earn a higher salary and have a higher net worth.

These conclusions on how the different predictors affect the final predicted purchase amount could sound so obvious. In fact, they are and it would not be necessary to be a genius in order to figure it out. The point of this study is finding and quantifying the actual relation and measuring the influence of the predictors in the model. This was shown to be clearly satisfied.

## 4. References

[1] E. García-Portugués. *Notes for Predictive Modeling*. Version 5.9.6. ISBN 978-84-09-29679-8. 2021. URL: https://bookdown.org/egarpor/PM-UC3M/.

[2] Dev Sharma. *Car Purchasing Model*. Apr. 30, 2021. URL: https://www.kaggle.com/dev0914sharma/car-purchasing-model (visited on 01/14/2022).