Classification model predicting High Value Properties

Damian Trzcinski

2022-06-05

Contents

List of Figures

1 Introduction

The intention of this project report is to present my findings, based on the analysis of the chosen data set - Real Estate Sales - with the techniques discussed in the course DS805. The ultimate objective of the project was to develop a good classification model, based on the data set provided.

Below is the list of assumptions I'm making when working with the assignment:

- As the assignment asked for building a Classification model, I understand that the variable that we want to predict (the response variable) needs to be a categorical variable
- By "selection of the optimal classification" rule I understand the rule the model uses to predict the class variable of the response variable, meaning the set or subset of explanatory variables that go into the model and are used for building predictions
- By "an additional classification rule" I understand a rule the model uses to predict the class variable of the response variable, meaning the set or subset of explanatory variables that go into the model and are used for building, different than the first Classification Rule proposed and by "classifier" I understand the Classification Rule of choice
- I assume that Data Transformation also includes selection of subsets of variables of the dataset, in order to focus the analysis on the best potential predictors and not all possible predictors and all possible combinations
- The report is based on the theory taught in the course DS805 and all course materials, including presentations and R code with custom formulas

2 Summary of the data

Dataset "Real Estate Sales" was a dataset provided in the assignment with a brief description of the data. As presented in that description, it consists of 522 observations on the transactions for the residential home sales prices in the mid-western city, obtained for home sales during year 2002. Each observation (line in the dataset) has been assigned with Identification Number and provides information on 12 other variables. After loading the data into R, we can observe that all the variables are non-negative integers, as they only take on values equal or higher than 0, without decimals. The dataset in R was called *realestate* and this is the name I am going to refer to "Real Estate Sales" dataset from now on.

Even though values for all variables are integers, it does not mean all variables are just numerical variables. According to my assessment, we can observe 3 types of variables in the *realestate* dataset - numerical discrete, categorical nominal and categorical ordinal. In the analyzed dataset we do not observe **numerical continuous variables**, as all the variables are integers, hence do not have any fractions. It is worth to mention though, that in the real world the variables *2. Sales price, 3. Finished square feet, 7. Garage size* and *12. Lot size* could be considered as numerical continuous variables as they represent measures, which are continuous and dependent on the setup precision of the decimals. In the brackets I provided the names of the variables in the R code, attached to this report as an Appendix.

The **numerical discrete** variables are the variables that can take any values within a finite or infinite interval, they must be integers and the values can be compared and ordered.

In the dataset *realestate* these are:

- 1. Identification number ("id_num")
- 2. Sales price ("price")
- 3. Finished square feet ("area")

- 4. Number of bedrooms ("bedrooms")
- 5. Number of bathrooms ("bathrooms")
- 7. Garage size ("garage")
- 8. Year built ("year")
- 12. Lot size ("lot")

The **categorical nominal** variables are the variables that can take limited number of possible values and their values cannot be compared or ordered.

In the dataset *realestate* these are:

- 6. Air conditioning ("aircon"), as it states weather a house has or does not have air conditioning
- 8. Pool ("pool") - same principle as for Air Conditioning
- 11. Style ("style") - as it defines a style of a residence, could be decoded into text and indicates that a house belongs to a certain category of houses. We do not know weather one category is better than the other
- 13. Adjacent to highway ("highway")- same principle as for 6. and 8

The last data type is the **categorical ordinal** and it is similar to the categorical variable, but there is a clear ordering of the values that the variable can take.

In case of the *realestate* dataset, there is only one ordinal variable - 10. quality (*quality*). It is ordinal variable, as there is a clear ordering of the quality scores - from 1 to 3, where 3 indicates low quality score and 1 indicates high quality score.

By running a *summary(realestate)* we can not only learn the basic information about the data, such as minimum and maximum values for the variables, mean, median and quantile distribution. Running that function will also allow us to identify if there are any missing values in any of the variables and present count of them, which would mean that we need to identify the missing values in the dataset and decide upon what to do with those observations.

There are no missing values in the dataset, therefore we do not need to action anything in this regard. However, if we found a few missing values, we could resolve it in various ways. We could potentially decide to exclude the observations with the missing values. Another approach would be to do the *imputation* of the missing data, based on certain assumptions, such as using mean instead of the missing values. If a number of missing values was significant for a particular variable but the remaining variables would not present the same proportion of missing values, we would potentially have considered excluding that variable from the analysis, as large number of missing values would influence accuracy and integrity of the analysis.

2.1 Intitial Data Transformation

Before proceeding with checking the necessary assumptions for conducting the analysis of the data, the loaded dataframe with 13 variables and 522 observations requires initial data transformation.

As we categorized some of the variables as categorical nominal and ordinal variables, there is now a mismatch between the loaded data types and the actual data types, as R did not recognize categorical variables correctly. Therefore, I applied *factor* function on the selected variables, to assign them to the right data types (lines 46-49,57 in the Appendix)

Furthermore, we need to keep in mind that the first variable in the *realestate* is an Identification Number and it's not a variable that should be considered when analyzing and building the model. This variable is only used to identify a particular observation in the dataset and does not explain any feature of the data in question.

2.2 Response and explanatory variables

According to the description of the dataset, attached to the "Description of the Final project for DS805", our target value (response variable) should be the column 2 - Sales price of a house in US dollars. The remaining columns (apart from the Identification Number - *id_num* column) will be considered as potential predictors - explanatory variables (11 in total). I have decided to follow that suggestion and focus on building a model that will be predicting the variable Sales Price, based on the number of explanatory variables. However, that would indicate that I will build a regression model.

I understand that the description of the Final Project asks for building a Classification model. Therefore, if I am to use Sales Price as a target value, I would need to transform it into a categorical variable. I have transformed variable *price* into a categorical variable *High Value Homes (high_value)*, which indicates whether a house is considered a high value house, which is the house above the Sales price of 300.000 USD. I have chosen that thereshold, as an arbitrary thereshold and it can be adjusted in the code. I have made a decision based on the 70th percentile of the Sales Prices in the dataset provided.

The transformed variable *high_value* will be the response variable of choice.

3 Checking necessary assumptions regarding the data

Before building the model, it is important to check necessary assumptions about the variables, as it would shape the way we will work with the data, select classification method and rules and also build and interpret the future model. That includes checking the association between the variables, checking if the variables are normally distributed, both in the univariate but also multivariate contex and finally checking the homogentity of covariance matrix.

3.1 Checking associations between variables

In order to check associations between variables, we might want to plot the variables against each other on a simple 2D scatter plot and do it for each pairs of the variables. While placing one variable on x axis, and second variable on the y axis, we will be able to verify if there is any apparent relation between selected 2 variables. The numeric variables can have different directions of associations:

- Positive - while variable x increases, variable y generally increases
- Negative - while variable x increases, variable y generally decreases
- No association - there is no clear relation between x and y variables, points seem to be scattered randomly on the plot or while x increases, y remains constant, etc.

We can also check associations between numeric and categorical variables, or 2 categorical variables but the interpretation can vary, depending on how many classes categorical variable defines. When considering association for the categorical variables, it is also important to consider the accumulation of variables per given class.

The associations between the variables can be easily plotted in R using *pairs* function on the dataframe. The function will return a set of scatter plots between each of the variables in the dataset. As we have 12 variables in total in the dataset (after excluding Identification Number), I have split the figure with scatter plots into 2 separate figures, where in each of the figures I have included the response variable *high_value* and a subset of explanatory variables. I have applied jitter function to a copy of the original dataframe to be able to see accumulation of observations, especially important for the categorical variables.

Based on the output of *pairs* we can observe that there seems to be a strong association between *high_value* and *area*, *high_value* and *year* and somewhat strong association between *high_value* and *highway/quality*.

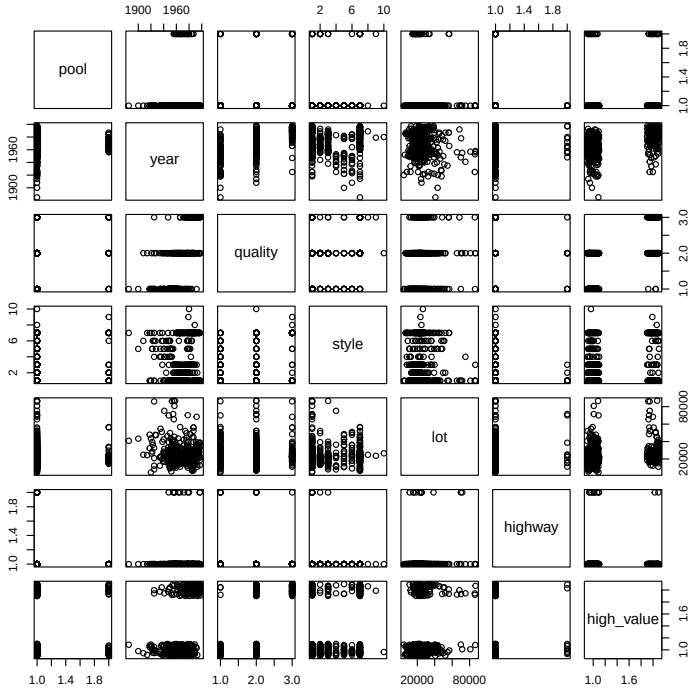

Figure 1: Associations between variables pool, year, quality, style, lot, highway and high_value

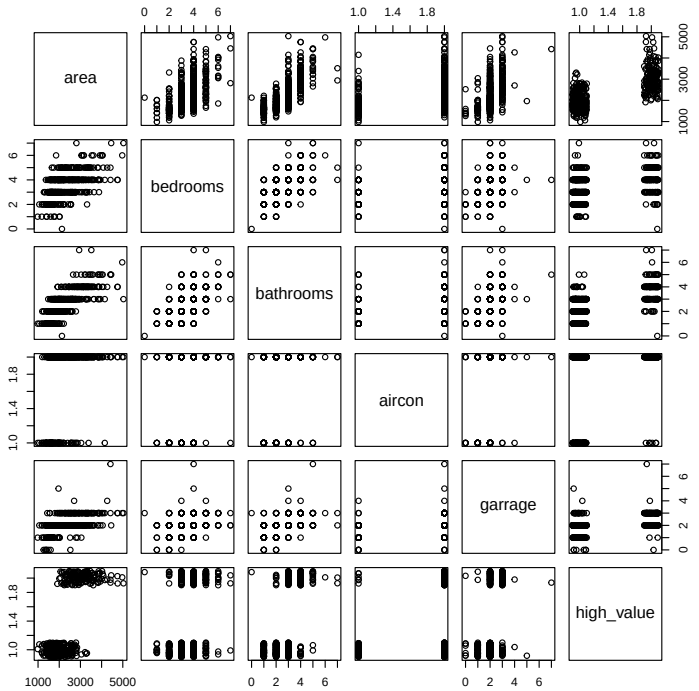

Figure 2: Associations between variables area, bedrooms, bathrooms, aircon, garrage and high_value

Even though in this step we are eyeballing the associations, this is one of the fastest and efficient ways to learn more about the behavior of the data in relation to the response variable, as well as to the other variables. When having a lot of variables, it can become really difficult to intepret the model, especially when trying to use all variables, so this step, where we are able to narrow down the data that seem to be interesting for further analysis and it's a common practice of Data Scientists in solving the real life problems.

Having mentioned the associations between the explanatory variables, we can also learn more about the data and make some initial assumptions. For instance, the scatter plots of pairs of variables *area* and *bathrooms* & *area* and *bedrooms* & *bedroom* and *bathroom* seem to be positively correlated in a similar way. This can mean that including all 3 of them might be redundant as the model containing all 3 might not be much better from the model that would be consisting only one of them.

For the numerical variables we are able to calculate covariance and correlation matrices, as well as assess if the variables are normally distributed and if the covariance matrices are homogeneous.

The covariance matrix is a square matrix that contains covariance values between the variable pairs. We have calculated the covariance matrix for all the numerical explanatory variables - *area, bedrooms, bathrooms, garage, year, lot.* Covariance matrix is sometimes difficult to analyze, as the units of measures for each of the variables included in the calculation of it is not standardized - meaning if there are huge differences in the scale of the values the variables present, the covariances can be really big number or really small numbers. The covariance matrix can for the most of it tell us if the relationship between the 2 variables is positive, negative or equal to 0. Only one relationship in the covariance matrix was negative - the relationship between *year* and *lot*, which means there is a negative correlation between these two variables.

In order to interpret the covariance matrix easier, we can compute the correlation matrix, which is basically the covariance matrix with the standardized units of measure. The covariance matrix can take on any numbers between minus infinite to plus infinite, whereas correlation matrix can take on values between -1 and 1, which makes it easier to compare the strength of association/correlation between the variables. Thanks to the correlation matrix we know, that the negative relationship between *lot* and *year* is very weak as the correlation for these 2 variables equals -0.10. This can be interpreted as one of the variables - for example year, explains only 10% of the variance happening in *lot* variable. This confirms the weak associations of these two variables that we could observe in Figure 1 already.

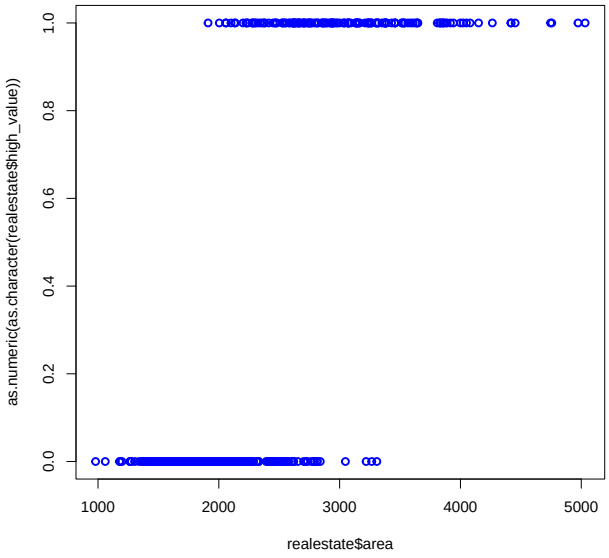

Figure 3: Scatter plot of association between variables area and high_value

The variable *area* seems to have the strongest correlation with the remaining numerical variables in the data set. This means that it explains for the most part the variance occuring in the remaining numerical variables and seems to be the strongest candidate variable for the future selection of variables for the classification rule.

If we select from the previously created pairs of associations for the variables in the dataset, only the variables *area* with the response variable *high_value*, we can see a clear split on the scatter plot between the two classes of houses and their area in square feet. We have previously assessed that association between variable *year* and *high_value* also looks promising, which means *year* might also be a good predictor, therefore I will also assess the necessary assumption for that variable.

As we have 11 variables in our data set, checking all the necessarily assumption of eg. normality and homogeneity for all variables and then pairs of 2 variables etc would be time consuming. In order to avoid unnecessary computations and be pragmatic, I will concentrate on the variables that are the most promising from the initial analysis of scatter plots, covariance and correlation matrices.

Based on that, I will focus on 4 variables: 2 numerical - *area, year*, and 2 categorical *quality and highway.* In the next section I will assess the normality and homogeneity of covariances for the selected numerical variables *area* and *year.*

However, after conducing the initial analysis and building the first model, I am planning to reevaluate if adding some of the eliminated from the initial analysis variables can improve the accuracy of predictions in the classification model in chapter 5.

3.2 Verifying normal distributions - univariate and multivariate

Verification if the variables are normally distributed is an important step in the process of understanding the dataset we are working with. The information on whether the data is normally distributed will then define if we would like to transform the data in order to potentially achieve the normal distribution, using for example Box Cox transformation (Johnson 2007, p. 193). It will also define what classification rule we can choose to build our future model, as some of the classification models require the data to be normally distributed, and others do not have that assumption.

For checking the univariate normality (normal distribution for a single variable) we are going to going to simulate the normal distribution for 522 observation - matching the number of observations in our dataset. Then, we will plot the theoretical quantiles from the normal distribution with the distribution of our data for a single variable and see how the distribution of the actual data matches the theoretical quantiles. That kind of plot is called *Q-Q plot.* Lastly, we will calculate the correlation between the simulated distribution and the actual data and assess, whether the actual data are normally distributed.

The null hypothesis H_0 in this case is - *Data for a variable x_n are normally distributed*

The alternative hypothesis H_1 would be - *Data for a variable x_n are NOT normally distributed*

We define our alpha as standard 0.05, which corresponds to the confidence interval of 95%. If obtained correlation is lower than the critical value for the given sample size and alpha value, we would reject the *null hypothesis* that the data are normally distributed. The critical value for the univariate distribution can be read directly from the table for Critical Points for the Q-Q Plot Correlation Coefficient Test for Normality (Johnson 2007, p. 181).

3.2.1 Area - Univariate Normal Distribution

Firstly I have assessed wheather variable *area* is normally distributed.

Based on the analysis conducted in R (line 145 of the code in the Appendix), we can already assume by looking at the Q-Q plot that the data are not normally distributed, as the actual data do not align with the constant line for the simulated data of the normal distribution. Calculation of the correlation seem to confirm this claim.

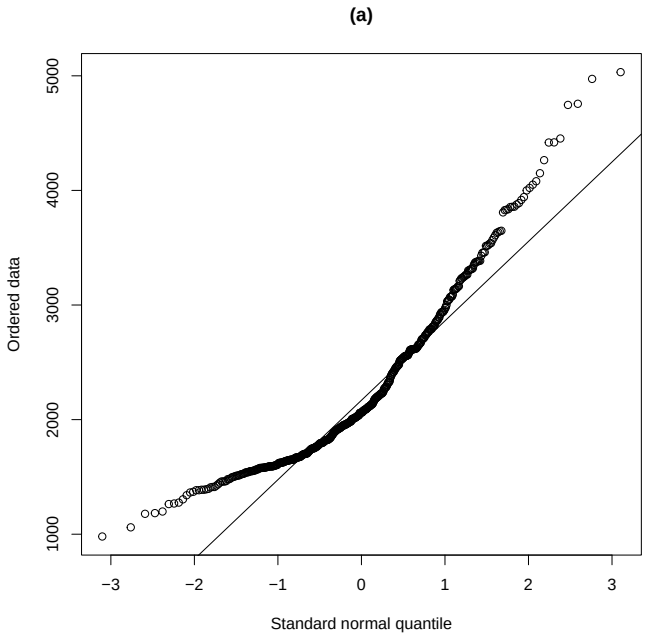

Figure 4: QQ Plot for the simulated quantiles from the normal distribution and variable *area*

According to the table of Critical Points for the Q-Q Plot Correlation Coefficient Test for Normality (Johnson 2007, p. 181), for the population 300 the critical value is 0.9953 for the confidence level 0.95 (alpha = 0.05), so for the population size 522 it would be even higher than 0.9953.

The correlation between the theoretical quantiles and ordered values of variable *area* is lower than the critical value ($0.9568 < 0.9953$), therefore we reject the null hypothesis that the data are normally distributed.

3.2.2 Year - Univariate Normal Distribution

Using the same method, I have assessed whether variable *year* has a univariate normal distribution. Surprisingly, only by looking at the Q-Q plot we can see that even though the data probably do not follow the normal distribution, they are closer to the normal distribution than the data for the variable *area*.

We are using the same critical value is 0.9953 as in the previous example and the confidence level 0.95 (alpha = 0.05) - it would be a value greather than 0.9953.

The correlation between the theoretical quantiles and ordered values of variable *year* is lower than the critical value ($0.9805 < 0.9953$),therefore we also reject the null hypothesis that the data are normally distributed.

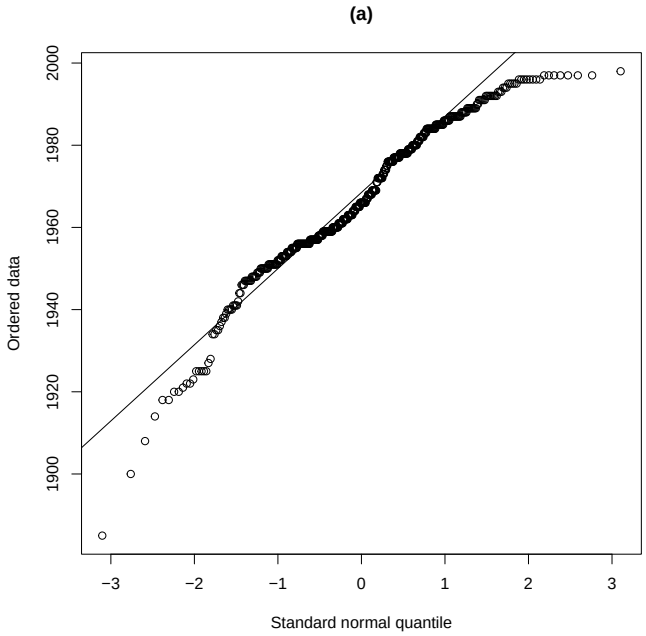

Figure 5: QQ Plot for the simulated quantiles from the normal distribution and variable *year*

3.2.3 Area & Year - Multivariate Normal Distribution

When we are assuming the Normal Distribution for 2 variables (bivariate normal distribution), we assume a chi-square distribution with 2 degrees of freedom (as we have 2 variables). If the data are in fact normally distributed, for relatively huge samples the observed Mahalanobis distance has an approximate chi-square distribution.

These two pieces of information can be used to assess whether some of the observations may be outliers and the data in question may have a multivariate normal distribution.

We are able to use the Q-Q plot to plot the Mahalanobis distance of the sample data. It starts with the same concept as the normal probability plot. For multivariate data - in our case we will consider 2 variables *area* and *year*, we need to plot the ordered Mahalanobis distances and estimated quantiles of samples of size 522 from the chi-square distribution of 2 degrees of freedom. As in the univariate normal distribution assessment, the data should match the straight line on the Q-Q plot. We will be also able to verify any potential outliers as they will be displayed in the top right corner of the plot. They will be located in that area of the plot, as the Mahalanobis distance will be significantly larger than the chi-square quantile value.

Having explained how the multivariate normal distribution of the data can be assessed for our selected subset of the dataset, we define the null hypothesis H_0 and alternative hypothesis H_1:

- H_0: The variables *area* and *year* have multivariate normal distribution for a sample size 522, 2 degrees of freedom and alpha = 0.05
- H_1: The variables *area* and *year* do not have multivariate normal distribution for a sample size 522, 2 degrees of freedom and alpha = 0.05

We need to find a critical value for chi-square distribution, that will allow us to make a decision weather we reject or fail to reject the null hypothesis. This can be done with the function *FindChiCrik* provided on the lectures.

The function *FindChiCrik* runs N times (1000 in our case) the following simulation:

- Generates n observations (522 in our case) from Chi-square distribution with p degrees of freedom (2 in our case) and sorts them in an ascending order
- Generates theoretical quantiles for the same parameters
- Calculates correlation between the 2 sequences generated above and saves it into a vector of correlations

After generating the correlation vector for N experiments it sorts it and returns a correlation on a position *ceiling(*$N \cdot$ *alpha)* as the critical value for our hypothesis test.

The function for specified parameters returned the value of 0.9909. This will be our decision value threshold - if a correlation between Chi-square quantiles and Ordered squared Mahalanobis distances will be lower than this value, we will reject the null hypothesis and assume the data are not normally distributed.

Firstly, I have calculated the means of each variable - *area* and *year* and marked it on the scatter plot below.

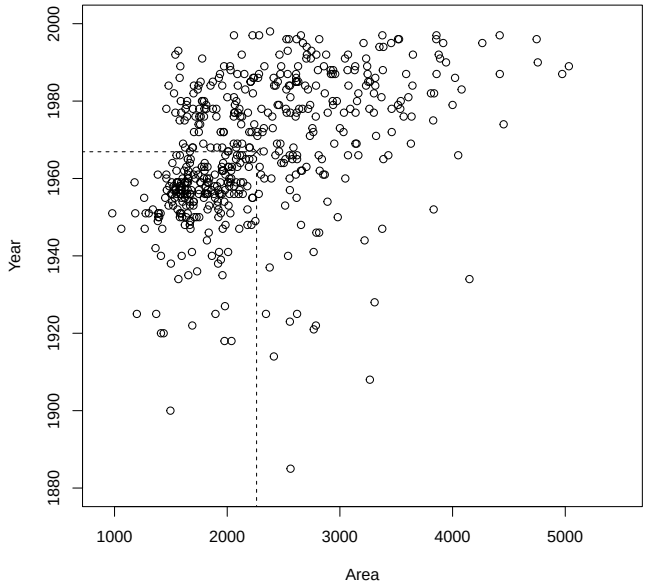

Figure 6: Scatter plot of the 2 variables in question - *area* and *year*

Then I have calculated the covariance matrix for *area* and *year* and inverted as inverted covariance matrix is needed for the calculate quadratic form that is then used to generate the *Contour Plot* for the Squared Mahalanobis distance. The Contour Plot is presented below:

Now we are able to evaluate bivariate normality. The first thing we are going to do is to count the number of points in the 0.25, 0.50 and 0.75 probability ellipse. One of the characteristics of the normally distributed data would be to have an expected proportion of data in each ellipse. We do that by calculating the squared Mahalanobis distance and comparing it with a simple proportion of 25%, 50% and 75% of the 522 observations in the data set. The expected vs the actual observation count is different for the 50th and 75th percentile, which could indicate the data is not normally distributed. The code can be found in the Appendix with the code (line 232).

This is only the empirical reasoning and we need to use Q-Q plot to verify it. We are able to evaluate bivariate normality, using Q-Q plot by plotting of squared Mahalanobis distances for the actual data against theoretical Chi-square quantiles. This in R will allow us to fetch necessary data for the calcuation of the correlation between these two variables and compare it with the critical value calculated earlier.

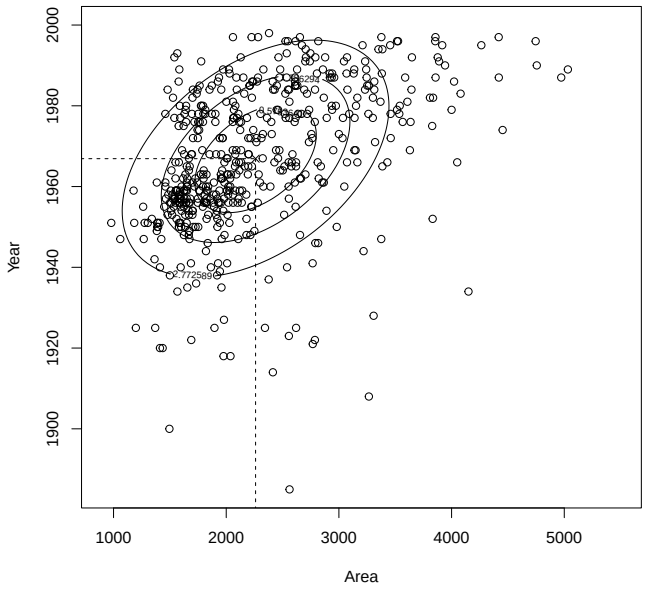

Figure 7: Contour Plot of the bivariate distribution of variables *area* and *year*. Ellipses represent the critical values for the Chi-quare distribution

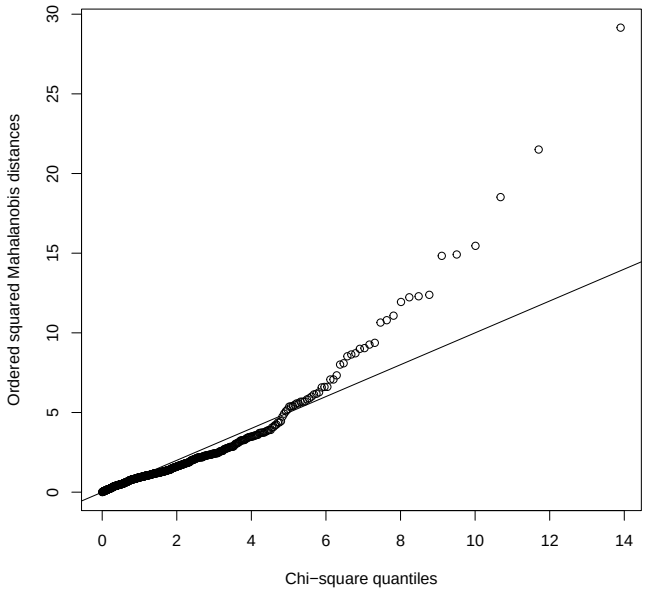

Figure 8: QQ Plot for the simulated quantiles from the Chi-square distribution and ordered squared Mahalanobis distances for variables *area & year*

As the correlation between Chi-square quantiles and the Ordered squared Mahalanobis distances for *year* and *area* is lower than the critical value(correlation - 0.9446 < 0.9915 - critical value) we reject the H_0 that the bivariate distribution for variables *year* and *area* is normal.

3.3 Data transformation - BoxCox

Neither Univariate nor Bivariate distributions are normal. I will now focus on verifying if Box Cox transformation will help to transform the Univariate distributions to normal distributions.

Box Cox transformation is a transformation that attempts to transform the data that are not normally distributed into normal shape (Johnson 2007, p. 193). Using the code provided on the lectures, I have transformed the univariate data, using the most optimal value of *lambda* and plotted it again against the theoretical quantiles as in previous sections, where we were assessing normality for the univariate variables without BoxCox transformation.

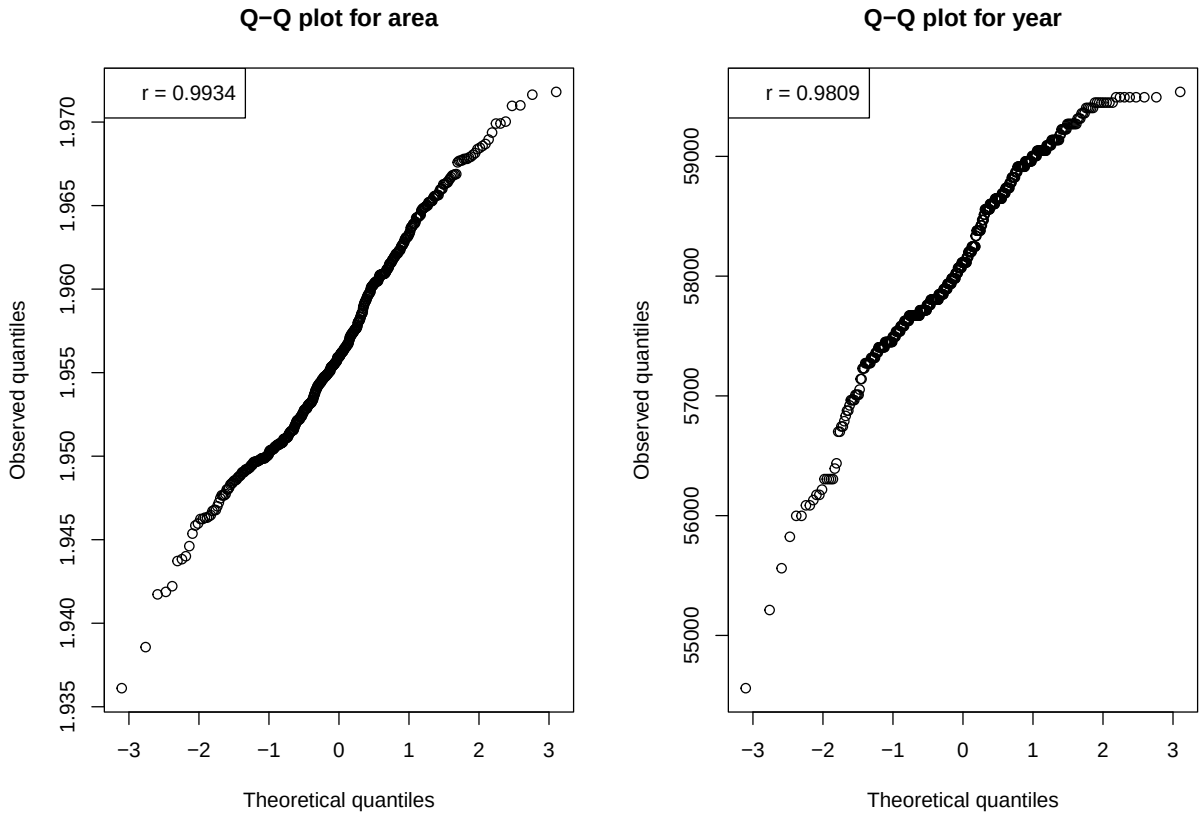

Figure 9: QQ Plot for the simulated normal distribution and transformed variable *area* (1st figure) and *year* (2nd figure) using BoxCox transformation

The correlation value from the Q-Q plot for the variable* *area* has improved significantly from 0.9568 to 0.9934 but it's still below the critical value of 0.9953, indicating that the variable is not normally distributed, considering critical value for alpha 0.05. The increase of correlation from Q-Q plot for variable *year* after Box Cox transformation is negligible.

As Box Cox transformations for Univariate distributions did not help us to achieve normal distribution for neither of the variable, in the next step I am going to check if removing outliers for the the bivariate distri-

bution of *area* and *year* helps to increase the correlation value between Simulated Chi-square distribution and the actual data, hence bring us closer to the normal distribution of those variables.

3.3.1 Data Transformation - removing outliers

Using code provided during the lectures, I have evaluted again the multivariate normal distribution of the variables *area* and *year*, but this time I have decided to remove the outliers and test, whether removing the last outlier helps to increase the correlation between the Simulated Chi-Square distribution and the observed Mahalanobis distance.

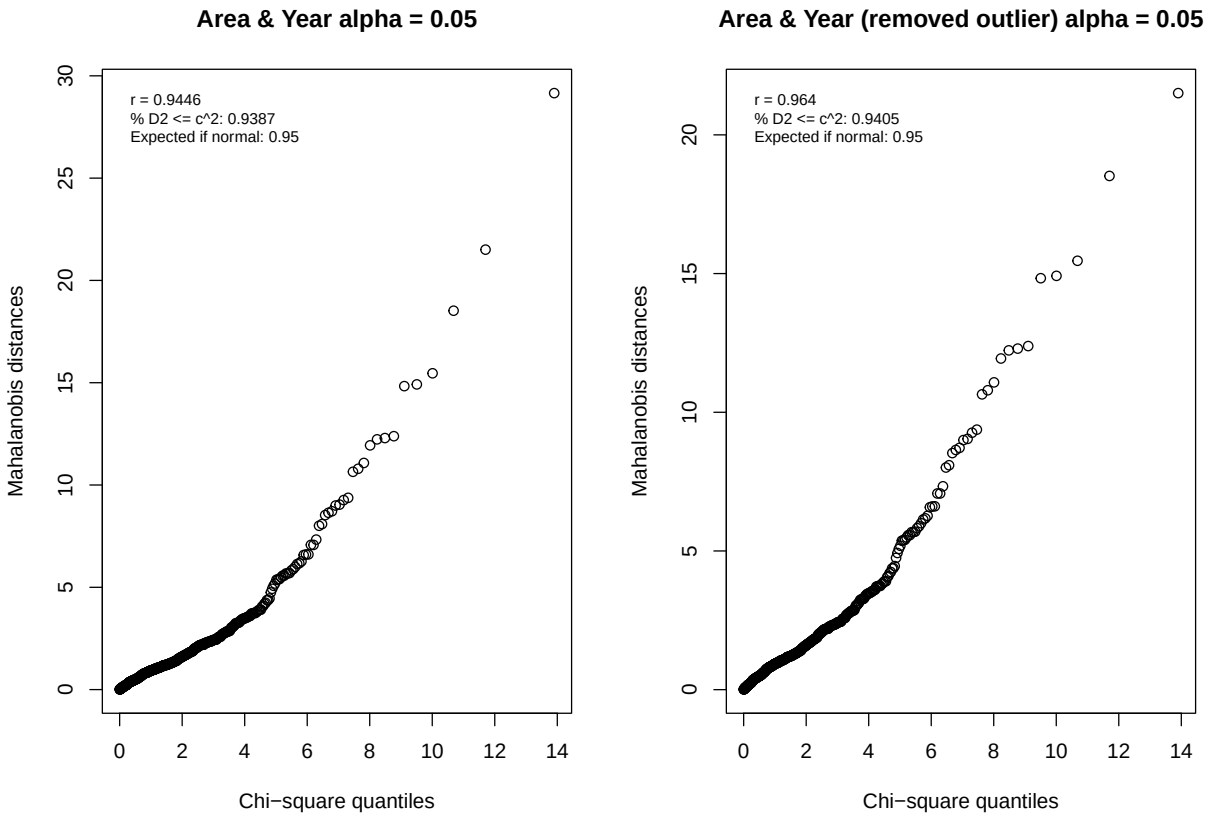

Figure 10: QQ Plot for the simulated Chi-square distribution and Observed Mahalanobis distance for variables *area* & *year*

Correlation of Q-Q plot for the bivariate distribution of variables *area* and *year* has improved when the outliers were removed. It still did not increase the correlation value above the level of the critical value which would allow us to assume normal distribution for this subset of variables from the dataset, but the correlation value increased from 0.9446 to 0.964, which is an improvement.

The data is still below the critical value of 0.9915 for multivariate distribution.

As neither Box Cox transformations of the single variables, nor elimination of the outliers for the bivariate distribution helped us to achieve the normal distribution of the data, I will use the original variables instead of the transformed ones, as using the transformed variables would increase the complexity of the interpretation of the future model, without providing us with the normally distributed attributes.

3.4 Verifying homogeneity of covariance matrices

Testing homogeneity of covariance matrices means an assessment of similarity of covariance matrices for the selected explanatory variables across the class labels of the response variable. We want to test if there is statistically significant difference between the covariance matrices for different class labels. It is an assumption that must be met when claiming the multivariate normality of the data (Johnson 2007, p. 310).

We are going to test homogeneity for the selected variables *area* and *year*, within 2 groups of the response variable *high_value*. This means that we have a set of observations belonging to class *below high value threshold (0)* and the second group belonging to the class *above high value threshold (1)*. We test homogeneity of covariance matrices for each of these groups, across 2 selected variables.

Using Box's M test in R - a test that is used to compare variation in multivariate samples (Johnson 2007, p. 311) - we calculate the covariance matrices for each of the groups, compute M value (M), Correction factor (u), Test statistics (C) and critical value (critvalue).

We define our null hypothesis H_0 and the alternative hypothesis H_1:

- H_0: The covariance matrices for the variables *area* & *year* are homogeneous between classes 0 and 1 of response variable *high_value* with confidence level 95
- H_1: The covariance matrices for the variables *area* & *year* are NOT homogeneous between classes 0 and 1 of response variable *high_value* with confidence level 95

As Test Statistics value (C = 57.05 is larger than the Critical Value (critvalue = 7.81) we reject the H_0, which means that covariance matrices are not homogeneous, between classes *high_value* = 0 and *high_value* = 1.

4 Selection of optimal classification rule

In this section I am going to present the selected Classification Model and Classification Rule. As the numeric variables in question do not meet the criteria for neither Univariate nor Multivariate normality, I have decided to go with the Logistic Regression as an appropriate classification algorithm, as logistics regression does not require the assumption of normality to build a model.

4.1 Building 1st model with 4 variables

Logistics regression is a supervised machine learning model used for Classification problems. It allows us to build a mathematical model, that predicts class labels for a categorical response variable. Under the hood the model calculates Log Odds (logarithm of odds). Below is the generic formula for the model:

$$Pr(Y = 1|X) = \frac{exp(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)}{1 + exp(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)}$$

However, in R it can be easily modeled using the *glm* function with the right parameters and probability can be calculated using *predict* function.

In the section 3.1 we have preliminary decided to proceed with 4 out of 11 explanatory variables from the dataset *realestate* in order to limit the size of this analysis and present a pragramatic approach to the model development. Variables to be fitted into the Linear Regression models will be *area*, *year*, *quality* and *highway*. Using those variables we are going to built a model that will attempt to predict the categorical response variable *high_value*.

Based on the output from the summary of the model (*summary(result)*) we can see that the only variables that are statistically significant are *area* and *year*, the remaining 2 variables are insignificant when *area* and *year* are present in the model. However, in this model I will proceed with including *highway* and *quality* in the model and in the next sections I will built models with different rule selection for further comparison.

Now we are able to run the model through our data and build predictions, based on the selected explanatory variables and present them in a confusion matrix. Confusion matrix allows us to feasibility visualize the performance of the Classification model, where we let the model to predict probability for the response variable - *high_value* in our case. We predict the *probability* and not the *Log Odds* as we setup in the *predict()* function parameter *type="response"*. Then we transform those probabilities into binary variables 0 and 1 and check if the model is predicting the correct class label or not, comparing it with the actual data using contingency table.

We also need to define the probability threshold, where the model will be classifying the class label as 0 or 1. In case of our model we have decided to go with the standard threshold of 0.5.

The confusion matrix presents the values split into True Negatives, True Positives, False Negatives and False Positives. In our case correct prediction of class label 0 / 1 and incorrect predictions.

	0 (predicted)	1 (predicted)
0 (actual)	349	18
1 (actual)	27	128

The model predicted 27 properties, which are *high_value* (0) as NOT *high_value* (1) and 18 properties that are NOT *high_value* (1) as *high_value* (0). Based on the data of predictions and actual data (labels), we check how ratio between True Positive Rate and False Positive Rate. The ratio will be changing, when we change a threshold used for the classification (in our case it was 0.5).

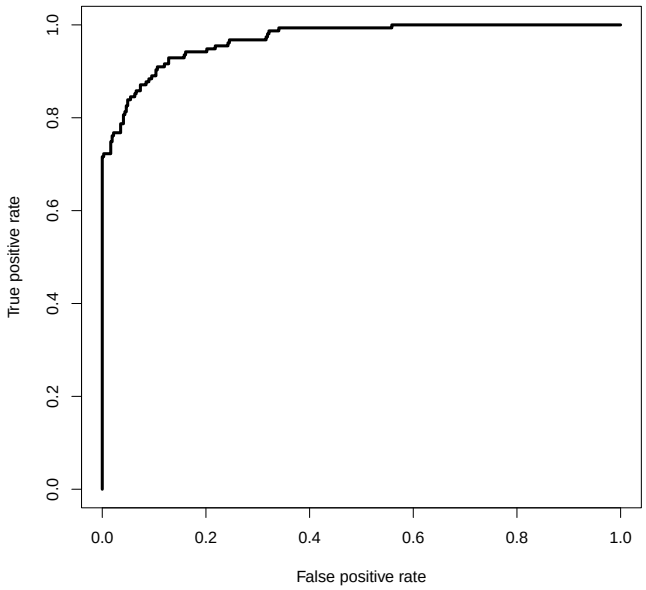

Figure 11: ROC curve for Model 1 with explanatory variables *area, year, highway, quality*

In order to calculate ratio between True Positive Rate and False Positive Rate we are going to use *Receiver Operating Characteristics* chart. It allows us to assess the performance of classification problems at multiple thresholds. AUC (Area under the curve) represents the degree / measure of separability and ROC is a

probability curve. It presents us with the information on how much the model is able of differentiate between classes. The higher the AUC measure, the better the model is performing at predicting 0 classes as 0 and 1 classes as 1.

Using functions from the *ROCR* library, we are able to draw the ROC curve on a plot and calculate the area under the curve, that is the performance measure of the model we intend to achieve.

Calculated measure of the area under curve is the probability of an event, where we present the model with 2 observations, one with class 1 and one with class 0 and the model will correctly classify them. In case of our first model that we have built, that probability (the area under the ROC curve) equals 0.9679. This means that 96.79% of the time the model will assign the correct class to the observations for the responser variable *high_value*, which seems to be a really good result.

4.2 Cross Validation of the 1st model

Cross Validation is a great method to verify if the model that we are building is not overfitting the data. When building a model (training), we present to the function all the observations we have, which means we do not have any test data left to verify if the model is reliable. Using Cross Validation allows us to shake the data a little bit by building a model using a subset of observations for the training phase and using the remaining data for testing the model.

In the functions presented during the lectures we have worked with Leave One Out Cross Validation, which means that we train the model on $n-1$ observations and test it on the remaining 1 observation left. We have able to iterate through our model n times and calculate the average performance of the model, considering all n interations. If the performance of the Cross Validated model is similar to the initial model, we can assume that our predictions with the initial model are reliable - the model is not overfitting to the current data.

Appendix with R code (line 520) presents the calculation for the Cross Validated model with 4 explanatory variables *area, year, highway, quality*. The confusion matrix for the model presents us with very similar results to the 1st model.

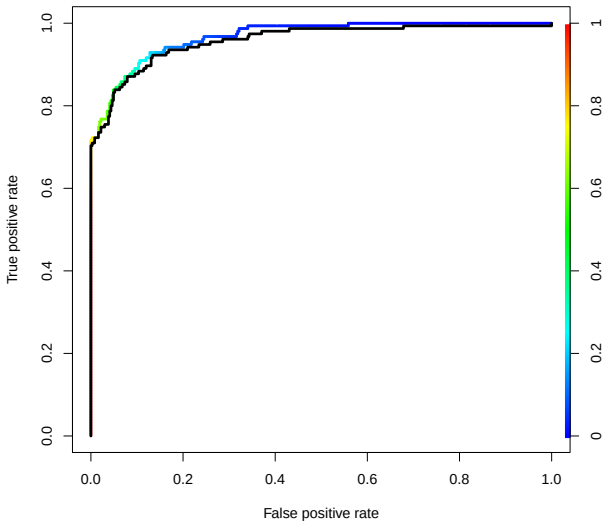

Figure 12: ROC curve for Model 1 with explanatory variables *area, year, highway, quality* (colorful curve) and for the same model, built with Cross Validation (black curve)

	0 (predicted)	1 (predicted)
0 (actual)	349	18
1 (actual)	29	126

We evaluate the performance of the cross validated model by plotting the ROC curve and calculating the area under the curve. The probability that the model will assign the correct class to the observations for the response variable *high_value* equals 0.9580 which is lower than the ROC of the initial model (0.9679). It is a comparable result and I have expected the Cross Validated result to be worse than the result of the initial model. The reason for that is that when building the CV model, we are not using the same pieces of information as we did for the 1st model, built without cross validation - we are using 1 observation less to train the model, therefore it is expected for the model to be less accurate. The difference is incremental, therefore we can say with high confidence that the model is not overfitting the data.

We need to keep in mind that we are working with *Leave One Out Cross Validation* which is not presenting the model with significantly different dataset - as we are excluding only one observation during the training. It is possible that the difference between the initial model and the Cross Validated model would be higher if we have decided to do 10 fold Cross Validation instead of 521 fold one.

5 Additional classification rules for further comparison

As we have previously worked with some of the data transformations, that we did not utilize for building the model (Box Cox tranformation of variables) as well as we have in total 11 explanatory variables and we only worked with 4 of them, I would like to verify if building Logistics Regression models with other parameters would give us different results than in the first built model in the section above.

5.1 Building 2nd model with variable *area* transformed with Box Cox

In section 3.3 we have tranformed variables *area* and *year* using Box Cox Transformation. Neither of the variables reached the thereshold for achieving the normality of the data distribution, therefore we have eliminated the idea of replacing the original variable with transformed one, as well as using transformed variable with the Box Cox requires additional data transformation to be able to interpret the coefficients of the model, which makes it more difficult.

However, even though variable *area* did not reach the threshold for it to be considered as normally distributed, after Box Cox transformation we have noticed significant improvement of the correlation between the transformed data and the theoretical quantiles. Therefore, for the sake of testing the effect of Box Cox transformation on the model accuraty, I have decided to built a second model with the transformed variable *area* included.

In the 1st model we have also used the variables *year*, *highway* and *quality* for building the model, but we have learnt, that only *year* was statistically significant for building a logistics regression model with our data, therefore in this model I will discard variables *highway* and *quality* from the model.

	0 (predicted)	1 (predicted)
0 (actual)	343	24
1 (actual)	26	129

We can see in the summary of the results for the model, that both transformed variable *area* and *year* are statistically significant for this model. The confusion matrix for this model gives us similar results as the 1st model from section 4.1. In order to verify that this model gives us similar accuracy as the first model, I have plotted the ROC and calculated the AUC, which equals 0.9630. As in the previous section, I have also calculated the Cross Validated model, which validates our result as AUC for LOCV 2nd model equals 0.9610.

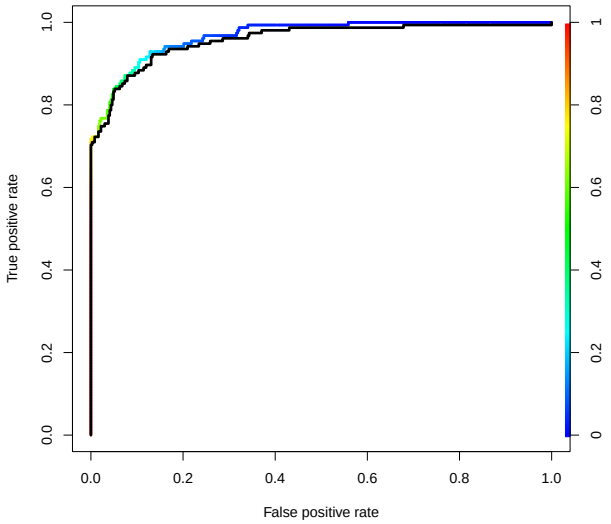

Figure 13: ROC curve for Model 2 with explanatory variables *area AFTER Box Cox tranformation, year* (colorful curve) and for the same model, built with Cross Validation (black curve)

5.2 Building 3rd model with all variables (except *style* and *id_num*)

In this subsection I would like to fit the model, using all explanatory variables from the dataset *realestate*. I assume it is possible to do so, as my selected Classification Model is Logistic Regression, which doesn't rely on the normality assumption, hence it is a good practice, but not a prerequisite for building Logistics Regression model.

I eliminate from the original DF column 1 - ID - as it is not a variable but an identification number for the observations. I will also eliminate variable *style* as there are some styles that occur only once, therefore it causes troubles when doing cross validation of the fitted model. Alternatively we could also remove problematic observations, but I would like to fit all models on the same number of observations so I decided to remove variable *style* from this model instead.

	0 (predicted)	1 (predicted)
0 (actual)	352	15
1 (actual)	23	132

The same process follows for this model as for the previous 2 models. We can see in the summary of the results for the model, variableS *area*, *year* but also *lot* are statistically significant for this model. This is a new information, as we did not consider the variable *lot* in our analysis.

The confusion matrix for this model gives us slightly better results, comparing to the 1st and 2nd model. But in order to quantify that difference, I have plotted the ROC and calculated the AUC, which equals 0.9773 and it's better by approx. 1,1 percentage point. As in the previous sections, I have also calculated the Cross Validated model, which validates our result as AUC for LOCV 3nd model equals 0.9641 and its close to the 3rd model, built without Cross Validation.

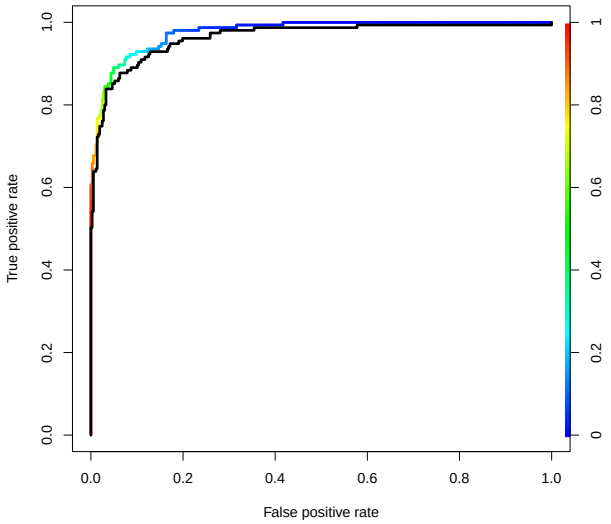

Figure 14: ROC curve for Model 3 with 10 explanatory variables (colorful curve) and for the same model, built with Cross Validation (black curve)

6 Evaluation summary of the classification rules proposed

In the previous section we have built models with 3 different classification rules and Cross Validated all of them to check if the results can be considered valid or if the models might have a tendency for overfitting. For all models we have calculated AUC, which can be found in the table below.

	Model 1 (4 var.)	Model 2 (2 var., *area* with BoxCox)	Model 3 (10 var.)
AUC	0.9679	0.9630	0.9772
AUC CV	0.9580	0.9610	0.9640

Looking purely on the AUC value, 3rd model presents the biggest accuracy. It is however the most complex one and contains many variables that are not statistically significant.

When it comes to the gap between AUC for the model and the AUC for the Cross Validated model, the smallest gap we can observe for the 2nd model. It can mean that this model has the lowest tendency to overfitting, which would make sense, considering there are only 2 variables used for building it and 1 of the variables was trasformed with the Box Cox transformation. The AUC value is not significantly different from the 1st model but interpreting coefficients in the model, where one of the variables was transformed with Box Cox transformation is not the easiest as it requires additional computations.

Even though we observe the increase in AUC in the 3rd model vs 1st model, adding all variables into the model does not significantly improve the accuracy of the model. At the same time it makes the interpretation of the model more difficult as there are more factors to consider and there are many variables that are not statistically significant. It puts the trained model at risk of overfitting, hence not being good enough for generalization of the predictions.

Having said that, it looks like from the models presented so far, 1st model is the simplest model from them all, includes mostly relevant variables, it's data are not transformed in any way, yet it provides a high accuracy of predictions, which means we would lean towards selecting classification rule similar to that model.

It is worth to mention, that we have learnt that the variable *lot* seems to be statistically significant as well, which we did not consider at the initial selection of the variables as this wasn't apparent. Fitting

more models, where we would be considering that variable in the model could be a potential next step in developing more accurate model for this dataset, but before we would do that, it would be a good idea to assess the normality of the data as univariate and multivariate distribution, which would incur additional computations, hence the cost. As I understand fitting more than 2 models is already beyond the scope of this assignment, I will not proceed with the analysis. Nonetheless, in the real world it could be a potential next step.

7 Selection of classifier

Based on the evaluation of the 3 fitted models I suggest to follow the Classification Rule proposed 1st model would be preferred, due to its high accuracy of predictions and relatively simple interpretation, as it has a low number of variables and none of the variables was transformed with Box Cox Transformation. However, during the process of fitting the model with 4 variables we have discovered that the 2 variables *quality* and *highway* were not statistically significant in contrary to high statistical significance of the variables *area* and *year*.

Therefore, I have selected even simpler model, consisting of only 2 statistically significant variables *area* and *year* as my prefered Classification Rule and built it as 4th (and last) model - results are attached in the Appendix with R code.

The confusion matrix and the ROC curve can be found below:

	0 (predicted)	1 (predicted)
0 (actual)	347	20
1 (actual)	27	128

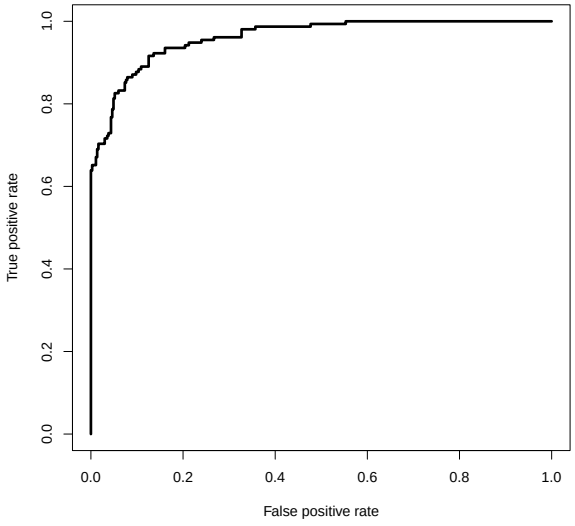

Figure 15: ROC curve for Model 4 with explanatory variables *area, year*

The final model is also presenting a high accuracy of prediction, with AUC value of 0.961. This means that 96.1% of the time the model will assign the correct class to the observations for the response variable *high_value* and it is using to predict it only 2 variables, which makes the model easy to interpret.

8 Conclusion

In this project report we have presented the analysis of the dataset *Real Estate Sales* and built a Classification model based on the analysis. We have presented evaluation of the properties of the data such as types of data, types of variables, calculated the basic statistics of the data and evaluated the normality of the variables, both univariate and multivariate for the selected subset of variables. We have also assessed the homogeneity of the covariance matrices for the selected subset of variables. Then we have selected the appropriate Classification method and tested 3 different Classification Rules in order to develop the final proposal for the Classification Rule for the formulated problem.

The Classification Model proposed is built using Logistics Regression as a method and it's using variables *area* and *year* from the original dataset to predict the class label for the variable *high_value* - the variable that evaluates wheather a house is worth more than 300.000 USD, thus considered a high value house.

The proposed Classification Rule is easy to understand, as it only uses two explanatory variables to predict the class label of the reponse variable. During the data analysis and evaluation process we have discovered, that using more than 2 variables does not significantly increase the model accuracy, evaluated with the ROC and AUC methods. We have achieved accuracy of the model on the level of 96.1% which is already a really good result. It can be considered a good result for the given problem, as we did not have a clear expectation towards what accuracy level we should achieve.

Transforming variable *area* with Box Cox transformation in order to achieve the Univariate Normal Distribution of that variable did bring us significance closer to the normality assumption but not fully. However, building the model with transformed data did not increase the accuracy of the model, therefore suggesting a model with that variable transformed using Box Cox would be redundant.

We have noticed that adding more variables into the model increased the accuracy of the model, but it was not a significant increase, therefore we suggested selection of a simpler model. When fitting the model to the data we have noticed, that potentially a next step when working with this data in the real world situation would be to also consider variable *lot* when building a model, as it showed potentiall to be statistically significant.

On the other hand we want to avoid the risk of overfitting of the model, which is not optimal, as it would negatively impact the ability of the model to generalize - predict class label of the response variable when a new observation is presented to the model - that is an observation that was not included in the training sample.

Potentially further transformation of the data could bring us even closer to higher model accuracy - for example, we could consider building PCA - Principle Component Analysis. PCA would allow us to start with more variables in the analysis of the relevant data for building a model, without increasing the complexity of the model. This can be achieved by using Principle Compontents that are a linear combinations of input variables and selecting the only the Principle Components that explain the most of variability of the model.

That approach could be beneficial, but being already on the accuracy level of more than 96%, we know it couldn't bring a significant improvement to the model. Additionally, it would be more computationally expensive, it would increase the complexity of the model, which would make it more difficult to interpret what pieces of data influence the final classification.

Analysis of the data and checking the initial assumptions is a huge step in building any model and for this classification model it was no different but it helped me to better understand how the data behaves and how to approach the process of building a model. In this case "less is more", as building a good classification model already two variables are doing quite a good job in predicting correctly the class label for the response variable.

9 References

JOHNSON, R. A., & WICHERN, D. W. (2007). Applied multivariate statistical analysis. Upper Saddle River, N.J., Pearson Prentice Hall.

QIN, J. (2022). Presentations to the lectures, Multiple publications, ItsLearning, DS805: Multivariate Statistical Analysis. SDU, Spring Semester 2022

QIN, J. (2022). R Code examples, Multiple publications, ItsLearning, DS805: Multivariate Statistical Analysis. SDU, Spring Semester 2022