Predicting Diamonds Price

Statistical Learning Mod. B, Final project

Bernardi Alberto, Cracco Gianmarco
23rd June, 2020

Università degli Studi di Padova

Goal of the project

Università degli Studi di Padova

Main Goal $\Rightarrow$ build an appropriate model to predict prices of diamonds.

In particular we will make use of:

1. Feature Selection;
2. Shrinkage Methods;
3. Principal Component Analysis.

Goal of the project

Main Goal $\Rightarrow$ build an appropriate model to predict prices of diamonds.

In particular we will make use of:

1. Feature Selection;
2. Shrinkage Methods;
3. Principal Component Analysis.

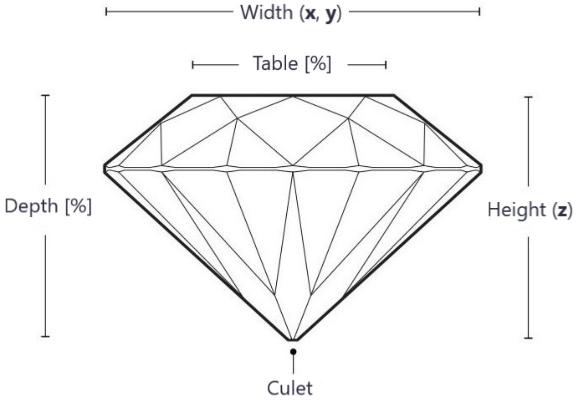
Obtaining the Data

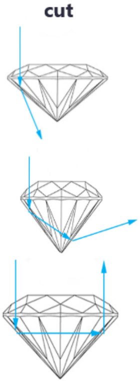
kaggle

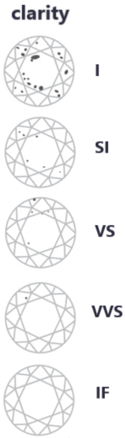
Diamonds Dataset from *kaggle*

https://www.kaggle.com/shivam2503/diamonds

Diamond Structure







Diamonds Dataset

Around 54000 diamonds with the following attributes:

- **Carat** (*numerical*): weight of a diamond (1 carat = 0.2 g);
- **Cut** (*factor*): *Fair*, *Good*, *Very Good*, *Excellent*, *Ideal*;
- **Color** (*factor*): *J*, *I*, *H*, *G*, *F*, *E*, *D*;
- **Clarity** (*factor*): *I1*,*SI2*, *SI1*, *VS2*, *VS1*, *VVS2*, *VVS1* and *IF*;
- **Depth** (*numerical*): percentual height over average width;
- **Table** (*numerical*): percentual width of the table over total width;
- **Price** [$] (*numerical*)
- **x**, **y**, **z** [mm] (*numerical*): three spatial dimensions.

Cleaning and preparing the data through the following:

- Check for Null Values;
- One-Hot-Encoding;
- Removal of absurd data;
- Check for possible outliers.

Data Preprocessing

Cleaning and preparing the data through the following:

- Check for Null Values;
- One-Hot-Encoding;
- Removal of absurd data;
- Check for possible outliers.

Data Preprocessing

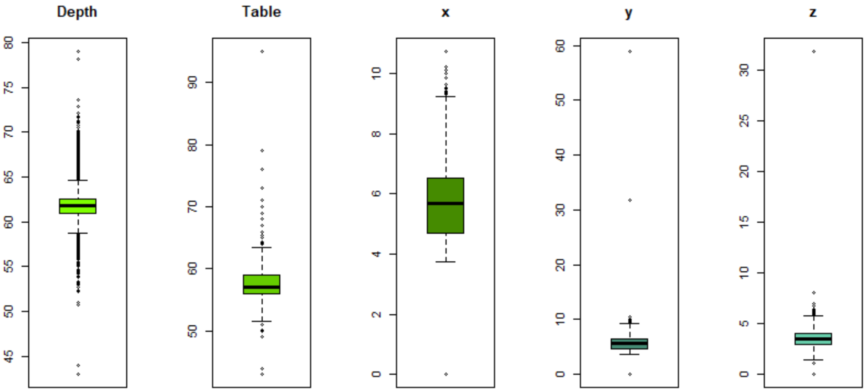
Cleaning and preparing the data through the following:

- Check for Null Values;
- One-Hot-Encoding;
- Removal of absurd data;
- Check for possible outliers.

Data Preprocessing

Cleaning and preparing the data through the following:

- Check for Null Values;
- One-Hot-Encoding;
- Removal of absurd data;
- Check for possible outliers.



For a more complete data visualization, take a look at the attached paper and **R** code.

Defining objectives

- Goal: Price prediction;
- How? Linear models ⇒ limited capacity, high interpretability;
- Why? Costumer advisory, market analysis, insurance quote.

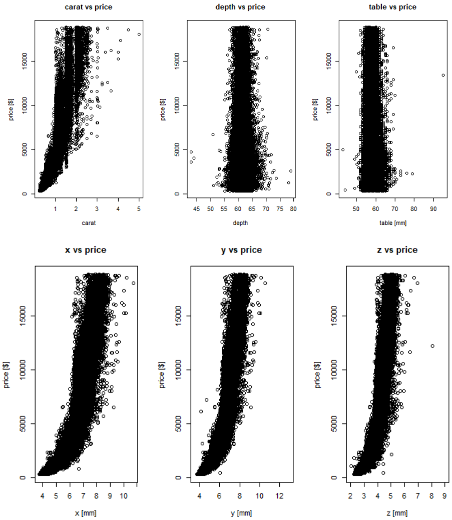
Defining objectives

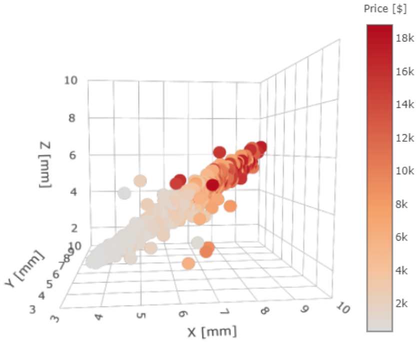
- Goal: Price prediction;
- How? Linear models $\Rightarrow$ limited capacity, high interpretability;
- Why? Costumer advisory, market analysis, insurance quote.

Defining objectives

- Goal: Price prediction;
- How? Linear models $\Rightarrow$ limited capacity, high interpretability;
- Why? Costumer advisory, market analysis, insurance quote.

Explorative - Price vs Numerical





Explorative - Price vs Categorical

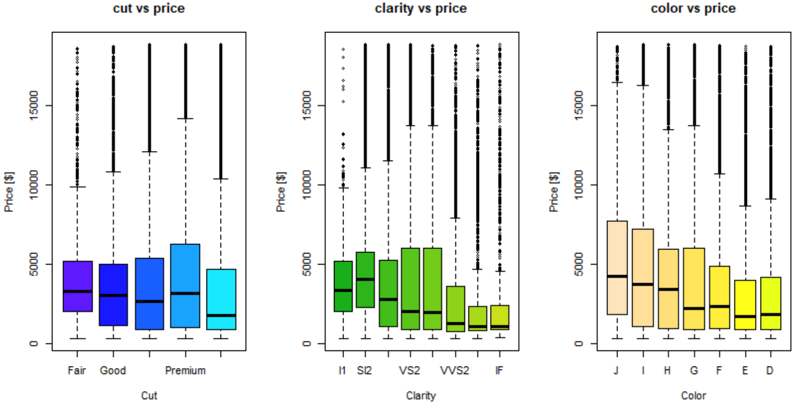


Figure: Boxplots of categorical features

Naive approach

First approach: we normalized our data (except for the target variable) and we tried to fit a linear model:

- Coefficients correctly describe the *expected* relationship between features and target variable;
- Analysis of residuals however shows they are clearly not randomly distributed, especially for the largest predicted values.

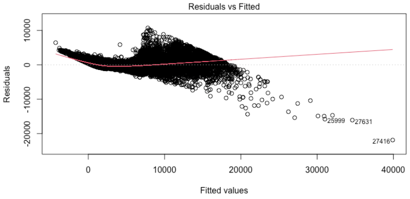


Figure: Plot of residuals for the naive approach

Naive approach

First approach: we normalized our data (except for the target variable) and we tried to fit a linear model:

- Coefficients correctly describe the *expected* relationship between features and target variable;
- Analysis of residuals however shows they are clearly not randomly distributed, especially for the largest predicted values.

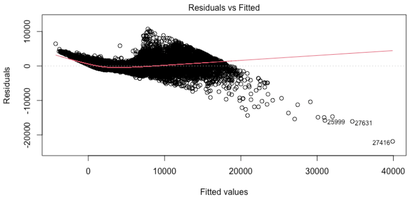


Figure: Plot of residuals for the naive approach

Solution: Variable transformation

To overcome this problem we followed two steps:

1. log(*price*): residuals become flat and uniformly distributed but now the coefficients are "wrong";

$\Downarrow$

2. log(*carat*): the coefficients are correct again.

Solution: Variable transformation

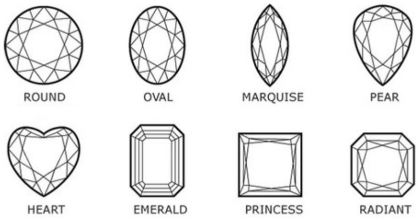
To overcome this problem we followed two steps:

1. log(*price*): residuals become flat and uniformly distributed but now the coefficients are "wrong";

$$\Downarrow$$

2. log(*carat*): the coefficients are correct again.

A new variable

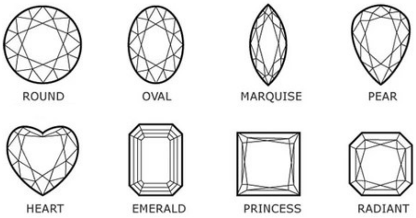


We introduce a new variable $r = \dfrac{x}{y}$

$$\Downarrow$$

Now we can capture the shape of the diamond.

A new variable



We introduce a new variable $r = \dfrac{x}{y}$

$$\Downarrow$$

Now we can capture the shape of the diamond.

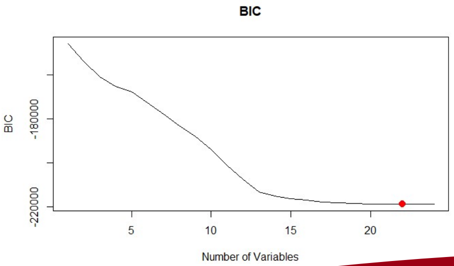
Best Subset Selection & Linear Model

We apply different techniques and criteria to select only the truly relevant features for our prediction.

$\Longrightarrow$ a simpler and more efficient model is obtained!

The final reduced linear model is given by:

$$\log(\text{price}) \sim \log(\text{carat}) + \text{cut} + \text{color} + \text{clarity} + x + y + r.$$

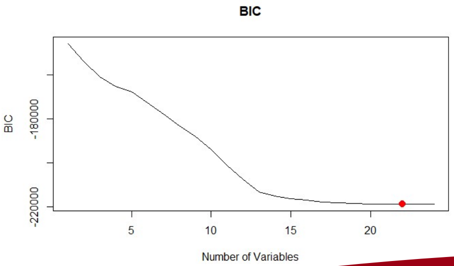


We apply different techniques and criteria to select only the truly relevant features for our prediction.

$\Longrightarrow$ a simpler and more efficient model is obtained!

The final reduced linear model is given by:

$$\log(\text{price}) \sim \log(\text{carat}) + \text{cut} + \text{color} + \text{clarity} + x + y + r.$$

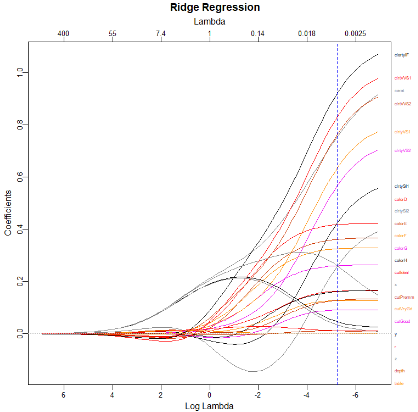


Shrinkage Methods

Other way to tackle feature reduction? Ridge regression and Lasso regression: two alternatives to Best Subset Selection in order to reduce model complexity.

$\Downarrow$

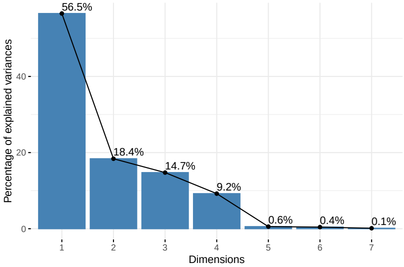
Features are not removed according to some metrics but instead the contribution of the irrelevant ones is shrinked toward zero.



One last method applied to obtain a model: PCA.

It is a technique that affects the data, not the model itself, and reduce their dimensionality.

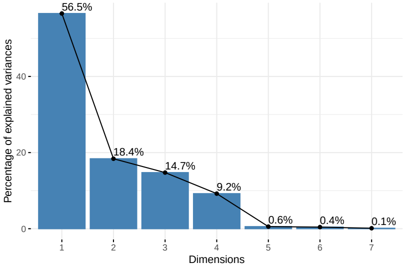
From 7 numerical features to 4 components!



One last method applied to obtain a model: PCA.

It is a technique that affects the data, not the model itself, and reduce their dimensionality.

From 7 numerical features to 4 components!

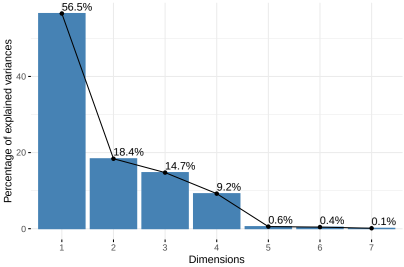


Principal Component Analysis

One last method applied to obtain a model: PCA.

It is a technique that affects the data, not the model itself, and reduce their dimensionality.

From 7 numerical features to 4 components!



Once PCA was performed ⇒ categorical variables have been added again.

Outcomes:

- Importance of categorical features;
- Importance of $\log(carat)$;
- Not enough dimensions to be really effective.

Once PCA was performed ⇒ categorical variables have been added again.

Outcomes:

- Importance of categorical features;
- Importance of $\log(carat)$;
- Not enough dimensions to be really effective.

Once PCA was performed ⇒ categorical variables have been added again.

Outcomes:

- Importance of categorical features;
- Importance of $\log(carat)$;
- Not enough dimensions to be really effective.

Principal Component Analysis

UNIVERSITÀ DEGLI STUDI DI PADOVA

Once PCA was performed $\Rightarrow$ categorical variables have been added again.

Outcomes:

- Importance of categorical features;
- Importance of $\log(carat)$;
- Not enough dimensions to be really effective.

Summary of what we obtained in this project:

	MAE [$]
Naive linear model	740
Linear model (reduced)	418
Ridge regression	472
Lasso regression	448
PCA	1245
PCA + cat. vars.	875
PCA with log(*carat*)	989
PCA with log(*carat*) + cat. vars.	564

Mean Absolute Error: the smaller, the better.

Further Development

In the attached report you can find way deeper insights with much more technical results.

Anyway, there is still much space for improvement!

- Non-Linear models;
- Machine Learning;
- Deep Learning.

Further Development

In the attached report you can find way deeper insights with much more technical results.

Anyway, there is still much space for improvement!

- Non-Linear models;
- Machine Learning;
- Deep Learning.