UNIVERSITE DE RENNES 1

YEAR: 2022-2023

PROJECT REPORT ON HOUSE'S PRICES PREDICTION

LEVEL : MASTER 1 MIAGE

STUDENT :

TOURE WILLIAM

Kaggle name: Luchagato

Kaggle team: LuchaLibre

TEACHERS :

MR MALINOWSKI SIMON

MR DELAUNAY JULIEN

I. PROBLEM

With several characteristics, we want to predict prices of potential houses. But, we are not in time series case as we do not depend on the time. So we will use predicting techniques based on linear and non-linear regressions. We have two datasets : houses for training and houses_competition for testing.

II. VARIABLES STUDY

Houses contains 20 variables but ID is not relevant as it's unique. We will use the others variables. The target variable is price as we want to predict it.

III. SIMPLE LINEAR REGRESSION

A simple linear regression is based on mathematics equation as $\mathbf{y = ax+b + e}$ where y is the target, a and b are coefficients, e is error and x is a simple predicting variable.

As we have 18 predicting variables, we want to know which variable is the best when it is alone. According to lecture, we have 3 criteria as Rsquared adjusted, Generalisation error and Student's critical probability. Let's see what do we get when we used them.

III.1.BASED ON RSQUARED ADJUSTED

$\mathbf{R^2_{adj}=(R^2(n-1)-p)/(n-p-1)}$ where n is the number of entry for y and p the number of variables. In our case, p=1. So $\mathbf{R^2_{adj}=R^2}$. As R^2 is not a good as criteria, we will not use it.

III.2.BASED ON GENERALISATION ERROR

Generalisation error can be estimated by the mean of all Mean Squared Error (MSE) of all models. As we are just creating one model, we are going to use MSE.

III.3.BASED ON STUDENT'S CRITICAL PROBABILITY

We need to find which variables have pvalues <0.05 and find the lowest one.

IV. MULTIPLE LINEAR REGRESSION

A multiple linear regression is based on mathematics equation as $\mathbf{y = b_0+\sum^p_{i=1}b_ix_i + e}$ where y is the target, b_i are coefficients, e is error and x_i are predicting variables.

We are going to use all the predicting variables. So we do not need to use any criteria.

V. MODEL BY VARIABLES SELECTION TECHNIQUE

According to algorithms of selection, we have criteria and performance evaluation. For algorithm, we are going to chose the best model. At the end, we will get only two model sinstead of 18 models. (We will compute Generalization error in the summary to let you know why. But in the code, you will already see it). **We did implement only forward selection technique**.

V.1. FORWARD SELECTION

V.1.a) Strict stopping criterion

- Rsquared adjusted: We use Rsquared adjusted as stopping criterion as soon as we don't find any better value.
- Gen error : We use Generalisation error as stopping criterion as soon as we don't find any better value.
- Pvalues : To select variables, we find the most significant variables.

V.1.b) Delta iteration criterion (For our case, I set δ =3)

- Rsquared adjusted: We use Rsquared adjusted as stopping criterion after δ iterations.
- Gen error : We use Gen Error as stopping criterion as stopping criterion after δ iterations.

V.1.c) Continuous criterion

- Rsquared adjusted: We use Rsquared adjusted criterion to find the best of all best rsquared adjusted based models.
- Gen error : We use Gen error criterion to find the best of all best gen errors based models. In order to do so, we had to split the train dataset in two datasets as train1 set and validation set.
- Pvalues : We use Pvalues criterion to find the best of all best pvalues based models. In order to do so, we selected all the models that contains significant variables. After that, we used Gen error on real train and test sets to find who is the best of the best.

VI. NON-LINEAR REGRESSION

For these models, we apply directly variables selections techniques. We will use generalization error as performance because its formula is more near to data than all other performance evaluators.

We did not implement exponential and logarithmic regression because of infinite values. Also, we could not change those values as there are important and very specific.

VI.1. POLYNOMIAL MODEL

For the all the criterion, we used the same principles as before. But all the datasets had to be updated to polynomial variables.

The best model is the one with the less generalization error.

VII. SUMMARY

MODELS	GENERALIZATION ERROR	KAGGLE ERROR
Simple regression	7.335738268177677	6.37662
Multiple regression	4.378727631485619	3.36559
Forward selection by Rsquared adjusted (We compared 3 models)	4.377557593684428	3.37337
Forward selection by Generalisation Error (We compared 3 models)	4.381554699896033	3.3766
Forward selection by Pvalues (We compared 2 models)	8.185269592346524	6.46667
Polynomial regression (We compared 3 models and it took at least 5 minutes)	3.3446754390631934	2.74081

We found that polynomial regression had the best result. We tried it with 2 to 5 powers. We decided to try it with 10 powers. We found a better gen error but Kaggle's one did think so. So we kept the powers from 2 to 5.