MPC Final TP : Predicting house prices

You will find in this document some instructions about this final TP. You are given a dataset (`houses.csv`) that contains data about houses that have been sold. For each house, you are given the sale price (our target variable) and some informations about the house (our predictive variables). There are 18 predictive variables to help you predict the sale price of houses. A quick description about the different variables is given on the Kaggle web page of this final TP :

In this final TP, you have to use the different prediction methods that we have seen during the course in order to design accurate models to predict the price of new houses given their descriptive informations (the 18 predictive variables).

You are also given another dataset `houses_competition.csv` that contains informations about 2365 new houses for which you are asked to predict the sale price using some of the best prediction models that you will design. You will use the Kaggle webpage of this TP to submitt your predictions. A score will be assigned to you for each submission. This score is based on the mean squared error of the predictions that you submitt. More informations about how to submitt your predictions on the Kaggle website are given a bit later in this document. Note that in the dataset `houses_competition.csv` you don't know the values of the sale price of houses, so you can not use this dataset to design and evaluate different prediction models. This dataset will only be used when you have selected good prediction models to predict the price of these new houses.

1 Expected work

1. A quick statistical study about the dataset : predictive varibales, target variable, link between the variables.
2. Simple regression models : you can use only one predictive variable from the 18. You are asked to choose the best one and evaluate its generalization error.
3. Multiple regression : use the 18 variables to predict the sale price. Estimate the generalization error of this model.
4. Variable selection : apply the different variable selection techniques seen during the course to select interesting models. Can you find better models than before ? (according to the estimation of the generalization error)
5. Non-linear models : add some non-linear variables (that you can couple with variable selection techniques) to try to design more accurate models.

During the TP, you have not implemented all the variable selection techniques shown during the course. Only the forward selection was asked in the TP and with a strict stopping criterion. For this final TP, you can extend your work with these different possibilities :

— implement the two other stopping criteria explained during the CM
— use the critical probability as a performance criteria for variable selection (for this performance criteria, only one particular stopping criterion can be used : stop when all selected variables are significant and no one from the other variables are significant when added). Significant means with a critical probability less than 0.05.
— implement the backward selection procedure (with different performance and stopping criteria)

You can also implement the cross-validation technique in order to better estimate the generalization error of a prediction model.

You are not asked to implement all of these extensions. If you correctly apply the different things seen during the TP, you will get a grade up to 14/20. Then, depending on the different extensions you implement, you can improve your grade.

2 Methodology

For each of the different models that you will design in the steps 2 to 5 above, you are expected to estimate the generalization error of the models. And then, if you find that a model seems to be interesting you can apply it to predict the competition dataset and get a score on Kaggle. Your overall results can be summarized into a table like this :

Model	Gen. Error	Kaggle error
Simple reg. : variable ?		
Multilple reg.		
...	...	...

The column Gen. error corresponds to the estimation of the generalization error of the models, and the column Kaggle error is the score that you will get on the Kaggle web page of this final TP. Note that you should try a model on Kaggle if it seems interesting (better than the previous ones that you have tried or quite similar, in terms of generalization error). In a real prediction setting, you would only have one attempt (with the model that seems the best). But here, you are allowed to try models as soon as it seems to improve what you obtained before.

3 Expected report

At the end of this final TP, you have to send me a pdf report explaining what you have done, together with a Jupyter notebook with your code. Deadline : **1st of May**.

4 Kaggle competition

In this section, we explain how to submitt your predictions on the Kaggle web page of this final TP.

1. Create an account on `kaggle.com`. It's free, you just need to indicate a valid e-mail address. Choose a username that allows us to identify you.
2. Go at the webpage of the final TP : `https://www.kaggle.com/competitions/project-2022-mpc-m1/`
3. Click on the 'Join Competition' button
4. To create a csv file containing your predictions (obtained with a given model :

```
# If you have one regression model named 'my_model'
pred = my_predictions(my_model, competition)
pred = pd.DataFrame({'ID':pred.index, 'Price':pred})
pred.to_csv('my_submission.csv', index=False)
# This will create a csv file 'my_submission.csv' with the predictions
# of the competition dataset
```

5. Click on 'Submit predictions' from the kaggle webpage of this TP.
6. Drag and drop your csv file into the 'Step 1' box. Wait a few seconds
7. Write a small description about these predictions (which model, which technique used) in the 'Step 2' box
8. Click on 'Make Submission'. Your score will be computed and written in the Leaderboard