EXAMINATION QUESTION PAPER - Course paper

Component of continuous assessment

GRA 65181
Data Science for Finance

Department of Finance

Start date:	10.11.2021	Time 09.00
Finish date:	17.11.2021	Time 12.00

Component weight:	50% of GRA 65181
Total no. of pages:	3 incl. front page
No. of attachments files to question paper:	2
To be answered:	In groups of 1 - 3 students.

Answer paper size:	max 20 pages excl. attachments
Max no. of answer paper attachment files:	2
Allowed answer paper file types:	pdf
Allowed answer paper attachment file types:	pdf, r, txt

Table of contents is optional. No bibliography is required.

Course Code: GRA 6518, Data Science for Finance
To be answered: groups of 1-3.
Instructions:

- You do NOT need to add a bibliography or reference list.
- You should upload two files: i) a pdf file, containing a short paper (max length 20 pages), ii) a file with your code, in any format (if you are having trouble uploading, convert to .txt or .pdf).
- Code can be in any language.
- *The paper should be readable by itself, without any reference to the code.*
- The paper does not need to be long, but try to use precise language and carefully document your choices and procedures.

Take-home exam, November 2021.

The files *y.csv* and *x.csv* cointain the well known *california housing* data:
y.csv: median house value (in a neighborhood; 1 is 100k usd).
x.csv: 8 features:

1. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
2. HouseAge: average house age
3. AvgRoom: average number of rooms
4. AvgBedr: average number of bedrooms
5. Population: population of the neighborhood
6. AvgOccup: average occupancy
7. Latitude: a higher value if farther north
8. Longitude: a higher value is farther west.

We can assume that all the data was collected at a single point in time (no time series structure).

Question 1 (20 points). Plot histograms of all the 9 variables and briefly comment on the aspects that seem relevant for building a good model.

Question 2 (20 points). It's most common in the machine learning literature to fit and predict y in this dataset without any transformation, that is, forecast the price. An alternative is to forecast and predict the log price. i) explain how predicting prices or log prices implies a different loss function in terms of how we may then want to use the prediction. (Hint: try to explain what we are predicting to someone who does not know what a logarithm is, but who will use the predictions), ii) asymptotically (if we had a very large sample), does it matter whether we fit prices or log prices if we fit a linear model by OLS? And if we fit a very flexible model by XGBoost?, iii) if you had to fit a linear model, how could you pre-process the data to improve its fit?

Question 3 (30 points). Use the first 15000 observations for training, and leave the rest as a test set to compare different models. If the models involve validation or cross-validation of some hyper-parameter, make sure to perform these by sub-dividing the training set, so that the model never sees the test set in the training phase.

a) Without using the features longitude and latitude, compute the root-mean-squared error (RMSE) of forecasts from: 1) OLS, 2) regression splines (GAM model), 3) bosting of regression trees. Be clear about which hyper-parameters are validated or cross-validated and how this is done, and on any modeling choices (you'll have to make some choices for regression splines and boosting). Aim to give sufficient details so that it should be possible for the reader to reproduce your results without looking at your code.

b) For regression splines, show plots of $f_i(x_i)$ for each of the six features and comment briefly.

Question 4 (30 points). Repeat a) from Question 3, but now using all the 8 features. Briefly comment on the results and on the role of longitude and latitude.