Final case study (#8)

Group project:

- Form 3 groups of 5 with your classmates to collaborate on the final case study.
- The dataset for this case study is called *case_8_games_training.csv and the objective is to produce predictions for the dataset case_8_cames_testing_masked.csv*.
- The file contain a daily dataset for a set of games in the month of September of 2016.
- The objective is to **predict the average playtime** per day.

1. Consider the following general guidelines:
   - Exploratory Data Analysis (EDA):
   - Perform EDA to gain insights into the dataset and understand its characteristics.
   - Analyze numerical and categorical variables using appropriate plots such as bar charts, pie charts, or frequency distributions.
   - Conduct correlation analysis to identify relationships between variables. Visualize the correlations using heatmaps or scatter plots.
   - Analyze the dependent variable (house prices) and explore its distribution and relationship with different covariates.

2. Data cleaning and transformation:
   - Clean the dataset by addressing possible missing values, outliers, and other data quality issues.
   - Convert categorical variables into dummy variables if necessary to make them suitable for modeling.

3. Model selection and training:
   - Adjust the data for training and testing.
   - Explore and compare different models suitable for predicting the outcome variable based on the given dataset.
   - Evaluate the models using appropriate metrics.
   - Select the model that performs the best based on the chosen evaluation metric.

4. Model tuning:
   - Fine-tune the selected model to optimize its performance.
   - Perform hyperparameter tuning using techniques like grid search with cross-validation to find the best combination of hyperparameters.

5. Final results:

- Send me your model's predictions by the 01st of July and I'll provide you with the missing average playtime in the testing data so you can calculate your fit in the testing data.
- Prepare a 20 minute presentation to discuss on Monday, 08th of July at 19h.
- Report on the performance of the best model using previously defined metrics.
- Evaluate the tuned model using diagnostic techniques. One extra point will be awarded to those who perform better. The game is on!

Remember to add comments to your code and, in your presentation, provide clear explanations for the choices made and present the findings and results of each task in a concise and understandable manner.

Suggestion: Use git to develop the project collaboratively.

Extra: Use docker to make the output into production. I.e., select the best model and make it available as a service to where the user introduces inputs (predictors, X) and receives as return the output (predicted price, Y). [see: https://medium.com/@sauravpattnaik2011/end-to-end-ml-pipeline-using-docker-fa4878abcc33]

Good luck with your case study!

Tips: Consider taking the log of variables. Consider creating new variables (e.g. share). Carefully decide on what to do with variables with lots of missings.