Machine learning for Finance

Individual project

House Price Prediction

This is an **individual** project due on the **9Th of February 2025 23:59** (This is latest date I can officially set an assessment deadline). I will accept to 16th of February. Assignments received after this point will incur penalties.

All projects should be completed using Jupyter notebooks which will be delivered at the end of the project.

Project Context

The real estate market is a dynamic and complex sector of the economy, influenced by numerous factors such as location, economic trends, and property characteristics. Accurate house price predictions are crucial for various stakeholders, including buyers, sellers, real estate agents, and financial institutions.

In this project, we aim to develop a machine learning model to predict house prices based on various features and try to understand the relationships we find.

The original data for Ames, Iowa (obtained directly from the Ames Assessor's Office) is used for tax assessment purposes but lends itself directly to the prediction of home selling prices. The type of information contained in the data is similar to what a typical home buyer would want to know before making a purchase and students should find most variables straightforward and understandable.

The dataset has provided some challenges for traditional economic methods. You can see a traditional attempt at modelling this data in the below reference:

Pardoe, I. (2008). Modeling home prices using realtor data. *Journal of Statistics Education*, *16*(2), 143. https://www.tandfonline.com/doi/full/10.1080/10691898.2008.11889569

We want to explore if we can do better with machine learning.

Project Task

Develop a regression model to try to predict the error made by the un-named companies house price estimate. The error is calculated by comparing the predicted price to the actual sale price at the time of the sale (the column “logerror” is the log of this error, you can try to predict error or logerror but make sure not to include one in the features). The model should aim to predict the price (or error in this case) at the time the property is listed, only data available at this time is allowable.

IMPORTANT AT ALL STAGES ADD COMMENTS IN NOTEBOOK TO DEMOSNSTRATE YOUR THINKING ON THE DECISION YOU MAKE. THIS IS IMPORTANT FOR YOUR GRADE AND IS BEST PRACTICE.

TASK A – DATA ENGINEERING

You will need to first explore, clean and transform the data and decide on the features to use. Below are some hints:

- House prices are financial data, which doesn’t tend to be normally distributed. Depending on the model you this might be a problem. Luckily, financial data is often log-normally distributed (this may also be the case for some of the features). Maybe you can control for this.
- The following hint is included with the dataset: An issue related to the intended use of the model, is the handling of outliers and unusual observations. We want to way our typical two concerns:
    - In general, it might not be a great idea to throw away data points simply because they do not match a priori expectations (or other data points).
    - Alternatively, if the purpose is to create a common use model to estimate a “typical” sale, it is in the modeler’s best interest to remove any observations that do not seem typical (such as foreclosures or family sales).

  There are at least 5 observations that are particularly important to find (a plot of SALE PRICE versus GR LIV AREA will indicate them quickly). Three of them are true outliers (Partial Sales that likely don’t represent actual market values) and two of them are simply unusual sales (very large houses priced relatively appropriately). **Justify your approach to removing any outliers.**

- Again we have lot's of features and lot's of categorical variables. As we have discussed some models will deal well with this others not. Make sure you have a reasonable final dataset for each model you will test in the next task. This mayu mean you sue different data for different models, remember we want the best model, so that means setting each model up for success with the data it needs. Maybe the more advanced feature selection and dimension reduction techniques we have found are useful for some models.

TASK B - PREDICTION

Build the overall "best" regression model to predict the house prices and report your best estimate of how you expect it to perform on new data in the market. Remember this assessment of performance is important for understanding the risk of making investments on this basis.

- You will need to decide how you define the metric and process to demonstrate which model is "best".
- Comment on any limitations this model might have when used by management for its desired purpose (house price prediction on new data).
- **Important** you will mostly be graded on your ability to apply the machine learning approaches we have learned in class properly appropriately. So please include the work that you test and not just the final best model.

TASK C – INFERENCE & MODEL UNDERSTANDING

Management is interested in understanding the model. You want to provide input on how the model makes it's predictions and how different features contribute to the estimate. If your best predictive model is difficult to find these relationships (perform inference) you can separately demonstrate what you know about how the model makes predictions and use a separate model to perform inference.

- Determine which features are the most important for predicting the target variable.
- Try to find and report on any interesting interactions between features.

- Discuss if these make sense and any information that you should provide management to help them interpret these results. Discussion and showing understanding is particularly important in this final task.

Deliverables

- **Important!** As you will not provide an explanatory video, please include as much information as possible in the jupyter notebook on what you are trying to do and why and how you interpret the results.
- You need to deliver all Jupyter notebooks and code used to complete the projects. You should make use of the markdown cells in the Jupyter notebooks to explain clearly what you are doing in each code cell and why. You should also use these cells to provide your answers to the questions such as showing which is your best classifier and how you decide it was best. Jupyter notebooks should be "solved" so I can see the results of running the code without having to run it myself.

Grading

- **Important** you will mostly be graded on your ability to apply the machine learning approaches we have learned in class properly. So please include the work that you test and not just the final best model. Try to apply as much of what we have learned as possible.
- You might receive extra points for demonstrating the ability to apply approaches not covered in the class.
- What your final model is (and how well it performs) matters much less than your ability to correctly apply different approaches, explore the problem, and understand the final results and model. Please try to demonstrate in the notebooks and the video.
- Well organized Jupyter notebooks and code will help ensure that your intentions are communicated and are likely to increase your grade.

For any questions please contact: iscott@novaims.unl.pt