Final Data Science Project

Project Overview

PA-VA Realty is a new company that is looking to break into the real estate markets of Pittsburgh and Richmond. The CEO of the PA-VA Realty is looking to understand what drives house prices in both Pittsburgh and Richmond. In other words, he is trying to understand what the most important factors in house value in each representative market. Similar to PA-VA Reality's competitors, e.g. Zillow, they are also looking to implement a model that predicts the price of a house, so that their team of realtors can easily identify over or under priced homes based on the comparison of the list price to the predicted house price.

Congratulations! You have been recently hired as a Data Science Consultant for PA-VA Realty. Your task, as outlined above, is to **predict housing prices** for homes in Pittsburgh and Richmond. In order to tackle this problem, you have been given a set of historical data of housing prices as well as detailed property information.

Two datasets provided include: *train.csv* and *test.csv*. The training dataset consists of 1,400 observations that include the price of the home and a number of property details (e.g. square footage, lot size, number of stories, etc.). The test dataset is the same except only contains 600 observations and does not include the price of the home. You will use this training data to build, implement and test models. You will then generate predictions based on the "best" model. The data dictionary on the last page of this document gives a description of each variable.

This project should be treated as a take-home exam and is to be completed completely independently. You may not consult with anyone about any aspect of the project (including even simple coding questions) other than the professor and TA.

The final data science project is due by **11:59 PM EST** on **Tuesday April 20.**

—— Absolutely no late projects will be accepted ——

Project Deliverables

1) **Predictions:** A single csv file with 600 test observations named "testing_predictions_XXX.csv" where XXX is your student ID number that contains three columns in the following order:
    - id: propertyId provided in the test dataset
    - price: predicted price of the home.
    - student_id: your student id number
2) **Technical Report:** A pdf report that outlines your process from start to finish in technical detail. Please name it "technical_report_XXX.pdf" where XXX is your student ID number. The report should NOT include any code. You may use at most 3 figures or tables. Limit of 4 pages (double spaced). This should (at least) touch on the following:
    - Introduction and description of exploratory data analysis.
    - Identification of Data oddities e.g. missing data, extreme values, etc. and how you handled them.
    - Summary of models considered. How many models seemed to perform "best" in terms of predictive accuracy? How'd you measure?
    - What were the most important variables? How did you measure variable importance? Were the variables deemed most important consistent across the top-performing models?
    - What were the most challenging aspects of this particular dataset? Were you able to mitigate these issues? Do you really trust your "best" model? If your job depended on this model, how worried would you be?
3) **Final (non-technical) Report:** Discuss your findings to a non-technical decision maker, in this case the CEO and realtors of PA-VA realty. Introduce your project and summarize a couple of your key findings from your research that could be useful to understand housing prices in Pittsburgh and Richmond. Limit 1.5 pages (double-spaced). "final_report_XXX.pdf" where XXX is your ID number. (Hint: Non-technical decision maker means that they will not know concepts such as: lasso, ridge regression, random forests, gradient boosted trees, tuning parameter, cross validation, mean squared error, bias, variance, overfitting, etc.)
4) **Code:** A single (or multiple) .R or .Rmd named "code_XXX.R" where XXX is your student Id. Your code should be thoroughly commented and able to be run from another machine provided necessary packages and data are loaded.

Rubric

1) **Accuracy (10%):** Model predictions on the test dataset will be graded based on Mean Squared Error. To earn full points, you simply need to have a lower test MSE than the base rate and a simple model. YOU DO NOT NEED THE BEST MODEL. You should not spend all your time trying to get the lowest MSE. This goal of this project isn't simply to find one amazing model – rather, it is to find strong model(s) obtained by correct reasoning and to understand what those variables imply as well as the uncertainty surrounding them. We simply include this, because if your model is not better than a baseline, why do you even need a model?
2) **Technical Report (50%):** The technical report will make up a significant chunk of your grade and should contain the guts of your process. The four main components of the technical report you will be graded on are:
    - *Introduction / EDA* – This should give an overview of the problem, general information of the data, identify data oddities, summary statistics, etc.
    - *Methods Overview/Details* - This should contain a summary of the methods explored and the various approaches that were considered.
    - *Summary of Results* - This should provide an overview of all of the results obtained. Comment on overall trends, contradictions between models, etc. You can include a table here if it helps summarize the findings
    - *Conclusions / Takeaways* - Based on the results described in the previous section, Summary of Results, describe what you feel can safely be concluded. If there are further tests/models that you think would be relevant to pursue given the overall results.
3) **Final (non-technical) Report (30%):** How would you explain the results to someone interested in your findings that doesn't have a statistics background? Discuss your project and findings in a non-technical manner. Identify and summarize at least 3 specific key takeaways from your work. These can include any useful and potentially actionable findings and/or specific aspects of the work that decision makers should keep in mind.
4) **Quality of Code (10%):** Does code run from another machine provided necessary packages and data loaded? Is code "readable" and well commented.

Data Dictionary

Variable	Type	Description
id	Character	Unique property identifier
price	Numeric	Price of the home
desc	Character	Description of home
numstories	Numeric	Number of stories
yearbuilt	Numeric	The year the home was built
exteriorfinish	Character	The exterior finish of the home
rooftype	Character	The material of the roof
basement	Numeric	Indicator of if the home has a basement
totalrooms	Numeric	Number of rooms in the house
bedrooms	Numeric	Number of bedrooms in the house
bathrooms	Numeric	Number of bathrooms in the house
fireplaces	Numeric	Number of fireplace in the house
sqft	Numeric	Square footage of the house
lotarea	Numeric	Lot area in square footage
state	Character	The state the house is located in
zipcode	Numeric	The zipcode the house is in
AvgIncome	Numeric	The average household Income in that zipcode.