Final Data Science Project

Project Overview

PA Reality is a new company that is looking to break into the real estate market of Pittsburgh. The CEO of the PA Reality is looking to understand what drives housing prices in Pittsburgh. In other words, he is trying to understand what factors are most important in determining the price of a Pittsburgh house. Similar to PA 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 Reality. Your task, as outlined above, is to **predict housing prices** for homes in Pittsburgh. 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 700 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 300 observations and does not include the price of the home. You will use the 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 independently. You may not consult with anyone about any aspect of the project other than the professor and TAs.

The Final Data Science project is due by **11:59 PM EST** on **Tuesday April 18.**

Project Deliverables

1) **Predictions:** A single csv file with 300 test observations named "testing_predictions_LAST_FIRST_ID.csv" where LAST is the student's last name, FIRST is the student's first name, and ID is the student's e-mail ID (see example below). The file should contain two labeled columns in the following order:
   - id: propertyId provided in the test dataset
   - price: predicted price of the home.

2) **Technical Report:** A pdf report that outlines your process from start to finish in technical detail. Please name it "technical_report_LAST_FIRST_ID.pdf" (see example below). The report should NOT include any code. You may use at most 4 figures or tables. Limit of 5 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 all models considered.
   - How many models seemed to perform "best" in terms of predictive accuracy? How did you measure this?
   - What were the most important variables? How did you measure variable importance?
   - 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? Is there other information you may want in order to improve the final model/predictions/recommendations further?

3) **Final (non-technical) Report:** Discuss your findings to a non-technical decision maker, in this case the CEO and realtors of PA reality. Introduce your project and summarize some key findings that could be useful to understand housing prices in Pittsburgh. Please name it "final_report_LAST_FIRST_ID.pdf" (see example below). Limit 1.5 pages (double-spaced). (Note: Non-technical decision maker means that they will not know phrases concepts such as but not limited to: lasso, ridge regression,

random forests, gradient boosted trees, tuning parameter, cross validation, mean squared error, bias, variance, overfitting, etc.)

4) **Code:** A single .R or .Rmd named "code_LAST_FIRST_ID.R" (see example below). Your code should be thoroughly commented and able to be run from another machine provided necessary packages and data are loaded.

*File names example: If the TA, Ryan Cecil, with Pitt e-mail RMC144@pitt.edu were to submit his files for submission for this project, they would be labeled as

1. "testing_predictions_Cecil_Ryan_RMC144.csv"
2. "technical_report_Cecil_Ryan_RMC144.pdf"
3. "final_report_Cecil_Ryan_RMC144.pdf"
4. "code_Cecil_Ryan_RMC144.R" or "code_Cecil_Ryan_RMC144.Rmd"

Rubric

1) **Accuracy (10%):** Model predictions on the test dataset will be graded based on Mean Squared Error. This is largely an "all or nothing" category -- to earn full points, you simply need to have a model with lower test MSE than a pre-established base rate. You do not need the best possible model available, and you should not spend all your time trying *ad hoc* things in search of the lowest possible MSE. This project isn't a "predictive competition" like you might find on kaggle.com. The goal of this project is to find strong model(s) obtained by correct reasoning and to understand what those variables imply as well as the uncertainty surrounding them.
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. Include test error estimates from the best overall model(s) as well as from the model(s) you ultimately chose to rely on.
    - *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, note this.
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 2 or 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
zipcode	Numeric	The zipcode the house is in
AvgIncome	Numeric	The average household Income in that zipcode.
Location	Character	Location of zipcode region with respect to Pittsburgh. Either in the city, partly in the city, or not in the city.
DistDowntown	Numeric	Approximate distance of household from downtown Pittsburgh.