Final Project Description

ISE – 364 / 464

In this final, you will fully analyze, clean, and engineer a dataset, followed by building what you consider to be the "best" model in order to predict a target variable.

In this dataset, you will be given a training set that consists of 1,385 rows and 71 columns, and a testing set that consists of 1,459 rows and 70 columns. The target variable that you are trying to predict in the training set is the 'SalePrice' feature and is what you will be trying to predict values for in the testing set. The remaining features consist of a variety of house related information and descriptions along with some residential descriptions and other miscellaneous information. In your report, you should include a discussion of what features you believe to be the most impactful in determining the selling price of a house. Since this is a regression task, naturally, you should be evaluating your model with the Mean Squared Error (MSE) score (or the Root-Mean Squared Error), as it is what will be used by the graders to evaluate your model's predictions that you will submit.

You will submit a .zip file that contains three things:

1 - An in-depth written report of your analysis, cleaning, engineering, and modeling that you performed on the dataset. This should include the 6 following sections (along with any relevant sub-sections): (1) Introduction and Initial Data Observations, (2) Numerical Data, (3) Categorical Data, (4) Final Engineering Tasks, (5) Modeling, and (6) Summary. You should write your report to an audience that has no understanding of data mining processes and do not know why you would do what you do. Thus, be very thorough in explaining your methodology, along with your reasoning as to why you make the choices that you do. You should include descriptions of all steps that you perform such as the following (to give a few examples): how you clean the data and why you do that, why you choose to feature engineer the features that you did, why you believe that those features will be good predictors, why you think your model is the "best" that you could obtain along with adequate proof that you have compared it to all the relevant models that you have learned about throughout the rest of this semester. You should include relevant plots to illustrate the points that you are making in your descriptions, but do not put hundreds of plots in your report... Visuals are meant to aid the reader in understanding a point you are trying to make. Thus, if you have a plot, you should have some description of what information you are able to gain from looking at it.

2 - A submissions.csv file that contains a single column of 'SalePrice' predictions for the testing dataset. These are the predictions that we will score your model's accuracy on and will dictate your rank with respect to the other groups model scores.

3 - All the code that was written and used to obtain your final predictions (remove all non-relevant code).

The majority of your grade for this project will be determined by the breadth of your knowledge and understanding of the full data mining process along with best practices.

(Extra-Credit Opportunities) There are two opportunities for extra credit associated with this project:

(1) If your group submits a .csv file (titled “submissions.csv”) of model predictions by November 24th at 11:59pm, everyone in your group will receive extra credit in the amount of adding 2 percentage points to your overall course grade.

(2) All groups will receive extra credit based on the best model submissions that are submitted by the project deadline. There are a total of 6 groups. As such, the group with the best performing model (in terms of smallest MSE score on the holdout testing set) will receive a total of 5 percentage points to their overall course grade. The second-best model will yield 4 percentage points, the third-best will yield 3 percentage points, and so forth. The team that has the worst MSE score on the holdout testing set will not receive any extra credit.