Statistical Methods: Winter 2020 Final

Project Report due: 11.59pm, Friday, March 20, 2020
Instruction: The project is individual and discussion among students regarding the project is not allowed. You can ask the Instructor or GTA if you have questions or ask questions using the Canvas Discussion board.

Project Data

The project data is given in the `R` data file `final_exam_dataset.RData`. The data is derived from Ames Housing Price Challenge in Kaggle but with some modifications, so, they are not the same dataset. However, information on the dataset can be learned from the Kaggle website.

The response variable for the dataset is `SalePrice`. There are 55 explanatory variables in the dataset - 28 numerical and 27 categorical variables. The description of the explanatory variables are given in the file `data_description.txt`. You may use the lecture slides of Week 10, `Week10_3.pdf`, to get a review on performing statistical analysis of a dataset for regression problems.

Project Goals

There are two main goals of the project -

1. **Explanatory Model:** Finding a regression model, which gives a relationship between the response variable `SalePrice` and the explanatory variables. Explanatory variables, that are *important for inference*, are `GrLiveArea, TotalBsmtSF, LotArea, GarageArea` and `Neighborhood`.

   Perform statistical inference procedures to test the relationship between `SalePrice` and the *explanatory variables of importance for inference*. The proposed model should have a minimum R^2 of 75% and contain at least 6 explanatory variables. The intent of this project is for the majority of your effort to be devoted to creating and reviewing this model.

   In order to achieve this goal, the may use the following directions -

   - You may eliminate any data points you deem unfit - such as outliers. You may use deletion and imputation methods to handle the problem of missing data. You may also drop explanatory variables to address the problem of multi-collinearity.

   - You may use transformations of response and explanatory variables to tackle the problems of non-constant variance and non-linearity in the model. But, taking interactions between explanatory variables is not advised.

- Perform model selection with explanatory variables that are *not important for inference.* After model selection, add the *explanatory variable that are important for inference* to the model and perform model checking and adjustment to get a proper model for inference.

- Perform statistical inference procedures to test the relationship between `SalePrice` and the *explanatory variable of importance for inference.*

2. **Predictive Model:** Finding a good predictive model, that gives a good predictive performance for the responses in the test data set `final_exam_test_dataset.RData`.

   In order to achieve this goal, the may use the following directions -

   - To create the predictive model based on the dataset, `final_exam_dataset.RData`, you may eliminate any data points you deem unfit - such as outliers. You may use deletion and imputation methods to handle the problem of missing data.

   - You may use transformations of response and explanatory variables to tackle the problems of non-constant variance and non-linearity in the model. But, taking interactions between explanatory variables is not advised.

   - Perform model selection with all the explanatory variables with a focus towards creating a model with good predictive performance.

   - Perform prediction of response `SalePrice` based on the selected model and the explanatory variable values in the test data set `final_exam_test_dataset.RData`. Find the root mean square (RMSE) prediction error on the test data set, that is, find

     $$\text{RMSE} = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (Y_i - \hat{Y}_i)^2} \quad \text{Note, here,} \;\; n_{test} = 120.$$

     Try to find a model with RMSE less than 5000.

Project Deliverables

The project report is due by Friday, March 20, 2020 at 11.59 pm. No late submissions will be accepted.

1. **Report** - Submit a report of at most four-pages (reasonable margins and font size) that summarizes the elements listed below.

   - **Introduction:** Project/data background and introduction to the problem.

   - **Model Building:** Dataset adjustment and initial model building using model selection.

   - **Model Checking:** Checking the model and dataset for problems and violations.

- **Final Model Formation:** Adjusting the model and the dataset to address the problems related to dataset and the fitted model.
- **Results on Inference and Prediction:** Results on the inferential questions and the prediction performance on the test data set.

2. **Support Files** -
   - A well-documented `R` code with which we could completely reproduce the content of your project report. **Please do not include the `R` codes within the project report**. But, instead create an Appendix with well-documented `R` codes.

Project Evaluation

The Project carries 35 points, but there are bonus 5 points, so the total achievable points is 30 points. The distribution of the marks will be:

- 10 points - initial data analysis, model building including model selection (**Introduction and Model Building**).
- 5 points - model checking and adjusting the dataset based on problems in the dataset (**Model Checking**).
- 10 points - adjusting the model based on problems with model assumptions (**Final Model Formation**).
- 5 points - inference and Prediction based on the fitted models (**Results on Inference and Prediction:**).
- 5 points - Project report presentation and writing.