INFSCI 510: Data Analytics

Assignment 2: Linear Regression

Note

Provide all justifications for your actions and the summary at the end right in your Jupyter notebook using markdown. Check out this markdown cheat sheet: https://www.markdownguide.org/cheat-sheet/

Tasks

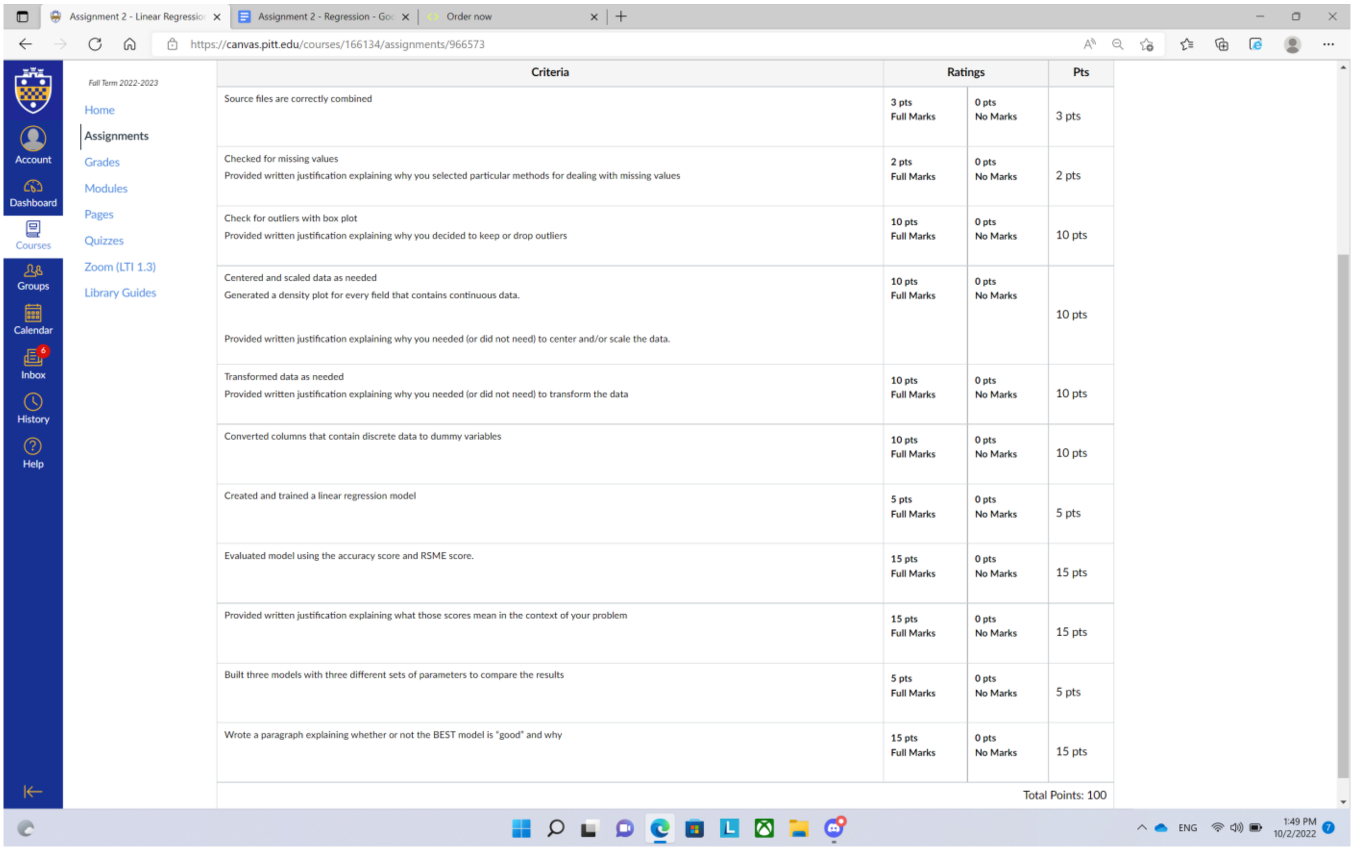
- Download a csv file containing real estate data data: https://drive.google.com/file/d/1hDWtr00dFzbeW4u4KEVPVUivu8ZksYKm/view?usp=sharing
- Using the data in this dataset, you will need to train a linear regression model. More specifically, you need to use a combination of house age, distance to the nearest MRT station, and number of convenience stores to estimate house price of unit area.

Below is a list of concrete tasks that you need to accomplish in your assignment

- Check for missing values. If there are any missing values, deal with them appropriately.
- Provide written justification explaining why you selected particular methods for dealing with missing values
- Check for outliers (Hint: box plot). Do we keep them or do we drop them? Why?
- Provide written justification explaining why you decided to keep or drop outliers.
- Center and scale data as needed
    - Generate a density plot for every field that contains continuous data
    - Review distributions
    - Chose centering and scaling approach
    - Provide written justification explaining why you needed (or did not need) to center and/or scale the data.
- Transform data as needed
    - Choose transformation approach
    - Provide written justification explaining why you needed (or did not need) to transform the data
- If there are columns that contain discrete variables, convert them to dummy variables
- Create and train a linear regression model that estimates the house price of unit area
- Evaluate your model using the R^2 score, adjusted R^2 score, and RSME score.
- Provide written explanations of what those scores mean in the context of your problem
- Play with predictors - will adding or removing predictors improve your model's accuracy? Build **three models with three different sets of parameters** to compare the results.

- Write a paragraph explaining whether or not your BEST model is “good” and why

Grading



Criteria	Ratings		Pts
Source files are correctly combined	3 pts Full Marks	0 pts No Marks	3 pts
Checked for missing values Provided written justification explaining why you selected particular methods for dealing with missing values	2 pts Full Marks	0 pts No Marks	2 pts
Check for outliers with box plot Provided written justification explaining why you decided to keep or drop outliers	10 pts Full Marks	0 pts No Marks	10 pts
Centered and scaled data as needed Generated a density plot for every field that contains continuous data. Provided written justification explaining why you needed (or did not need) to center and/or scale the data.	10 pts Full Marks	0 pts No Marks	10 pts
Transformed data as needed Provided written justification explaining why you needed (or did not need) to transform the data	10 pts Full Marks	0 pts No Marks	10 pts
Converted columns that contain discrete data to dummy variables	10 pts Full Marks	0 pts No Marks	10 pts
Created and trained a linear regression model	5 pts Full Marks	0 pts No Marks	5 pts
Evaluated model using the accuracy score and RSME score.	15 pts Full Marks	0 pts No Marks	15 pts
Provided written justification explaining what those scores mean in the context of your problem	15 pts Full Marks	0 pts No Marks	15 pts
Built three models with three different sets of parameters to compare the results	5 pts Full Marks	0 pts No Marks	5 pts
Wrote a paragraph explaining whether or not the BEST model is “good” and why	15 pts Full Marks	0 pts No Marks	15 pts
			Total Points: 100