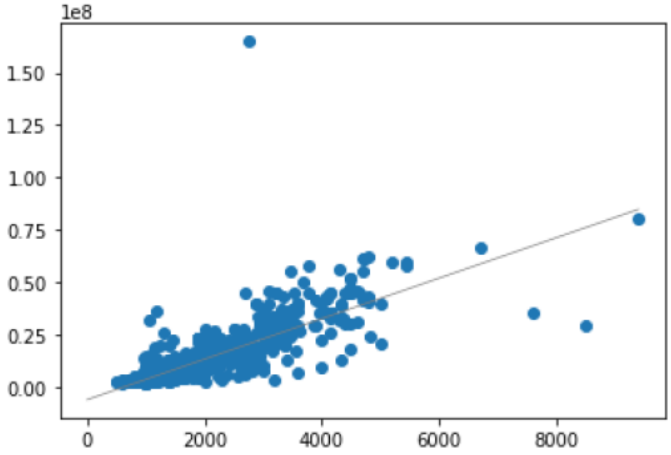
COMP-2704:
Supervised Machine Learning



Assignment 1: Linear Regression

Honour Statement

In doing this project, you must adhere to the following honour statement:

> *Red River College is committed to protecting the integrity of our curriculum ensuring the college continues to add value to our students and industry, while ensuring that students have opportunities to pursue their marks fairly, honestly and ethically.*
>
> *This includes, but is not limited to, the fact that no collaboration, plagiarism, cheating, unauthorized collaboration, or false representation is permitted on assessments and is a violation of the S4 – Academic Integrity policy.*
>
> *I understand I am subject to all of the same academic honesty requirements that apply during an in person assessment or online assessment. I understand that by beginning an exam, I accept and agree not to commit any violation of academic integrity.*
>
> *I understand that there are consequences for violating the policy and as a Red River College student, I will not participate in or condone academic dishonesty.*

Problem

1) *[5 marks]* Complete the following steps to setup this assignment.
   a) Open Jupyter Notebook in a web browser.
   b) Create a folder named "SupervisedML" and navigate into this folder.
   c) Within the "SupervisedML" folder, create a folder named "Assignment1" and navigate into this folder. The path to this folder should be "~/SupervisedML/Assignment1".
   d) Open the link: https://github.com/luisguiserrano/manning/tree/master/Chapter_3_Linear_Regression
   e) Download the files:
      - *House_price_predictions.ipynb*
      - *utils.py*
      - *Hyderabad.csv*
   f) Upload these files to the folder you created "~/SupervisedML/Assignment1".
   g) Open the *House_price_predictions.ipynb* notebook and run all code cells. Fix any errors that occur.

2) Use the *House_price_predictions.ipynb* notebook to answer the following questions. Insert text cells to write out your answers; be sure to state the question number.
   a) *[2 marks]* How many rows and columns are in the data file?
   b) *[2 marks]* Just from looking at the data, are there any potential outliers shown in the price vs. area scatter plot? List the *(x, y)* coordinates of any potential outliers.
   c) *[2 marks]* What is the equation of the best fit line produced by *simple_model* that uses only price and area?
   d) *[2 marks]* What coefficient values are associated with the intercept and area in the trained *model* that uses all features? Are these values the same as the intercept and slope found by *simple_model*?
   e) *[2 marks]* Using *model*, what is the predicted price of a house with three bedrooms and an area of 1000 square feet?
   f) *[3 marks]* What is the maximum error and root-mean-squared error of *model*? Explain what these mean.

3) Create a notebook with filename *SML_a1_q3.ipynb* within the folder "~/SupervisedML/Assignment1". Add cells with markdown text to the notebook to complete the following steps. Add text cells to answer the questions. You may copy relevant lines of code from *House_price_predictions.ipynb*.
   a) *[1 mark]* Import the necessary libraries and modules.
   b) *[1 mark]* Import the data from *Hyderabad.csv* into an SFrame named "data".
   c) *[2 marks]* Show two scatter plots: *Price* vs. *Area* and *Price* vs. *No. of Bedrooms*.
   d) *[4 marks]* Create a model called *two_feature_model* that uses *Price* as the target, and *Area* and *No. of Bedrooms* as features. Train the model and list the coefficients of the optimal solution.

e) *[2 marks]* Use your trained model to predict the price of a house with:

   - 6000 square feet and 4 bedrooms;
   - 1000 square feet and 3 bedrooms.

f) *[2 marks]* What is the maximum error and root-mean-squared error of *two_feature_model*? Compare these values with the errors for *model* and state which is better.

4) *[3 marks]* Upload your two notebooks to the Assignment 1 dropbox on the course website **before the due date** to complete your submission.

Total marks = 33