CS 457 - Homework Assignment 8: Regression

Due Date: Monday, October 30 at 11:59 pm

Purpose:
Demonstrate understating of Regression technique for correlation and prediction

Points: 100

Deliverables: Submit ipynb code file with your answers

- Review IDMA Book Chapter 12 Predictive Analytics
- Use the dataset `EmployeeSalaryRegression.csv`
- Perform analysis on the following questions**.** Make sure to include interpretation of each result including coefficients, p-values, r-square and other necessary information to support your answer

1. Create a regression model between TotalWorkingYears and MonthlyIncome (response variable). Show the scatter plot with regression line between them.
2. Create a regression model between Age and DistanceFromHome (response variable). Show the scatter plot with regression line between them.
3. Calculate Correlation for (1) and (2) and explain the values to support your answer.
4. Create a regression model to predict MonthlyIncome using all other inputs. Discuss the effectiveness of the model. Report 3 most significant inputs and 3 least significant inputs (based on p-value) and interpret the results. Create a new input and show the prediction of MonthlyIncome using the same model.
5. Create a regression model to predict HourlyRate using all other inputs. Discuss the effectiveness of the model. Report 3 most significant inputs and 3 least significant inputs (based on p-value) and interpret the results. Create a new input and show the prediction of HourlyRate using the same model.
6. Select only top 3 inputs (based on p-value) from (Question 4) and create a new model to predict MonthlyIncome. (If you are picking a categorical column for one of your top 3 columns, lets say JobRole, then make sure to add all the one hot encoded columns for JobRole and it will be counted as one top input). Discuss the performance of the model using few inputs as compared to using all inputs in (Question 4). Which model do you prefer and why?
    a. The idea is to create a simple generalized model with fewer inputs which are important for prediction. For this concept, please research and study "Regularization in Regression"
7. Using the model with 3 inputs in Question 6, create 2 new data records and predict their MonthlyIncome. Discuss if the predicted output makes sense based on your new data records.