COMPSCI 762 2022 S1 Week 6 Questions – Regression & Preprocessing (Part 2)

Luke Chang

March 27, 2022

Question 1

You are interested in the used car market. The year of manufacture and the asking price for a particular car model are given as the following:

Year	1985	1990	1991	2001	2004	2007	2007	2008	2008	2010	2013	2014	2015	2017	2018
Price	11800	6600	21800	4400	13000	14600	14700	15600	13200	17400	19200	8200	21500	27300	24700

1. Is there any outlier in the data? If it's, how do you determine outliers?
2. You realized cars that are in the same generation often asking for a similar price, so you deicide to use a binning method on **'year of manufacture'**. What are the results of smooth by bin means, smooth by bin median, and smooth by bin boundaries?
3. If you deicide to use **'year of manufacture'** to predict **'asking price'**, do you need to apply a data transformation technique before building the regression model? What are the available data transformation methods for this task, and what method do you think fit this scenario?
4. Build 2 regression models. The 1st one uses the original data. The 2nd one uses the preprocessed data. You may apply any preprocessing methods you think fits. Predict the asking price of a car manufactured in 2022. Which model do you think has better performance, and why? (You can use `sklearn` or compute it by hand.)

Question 2

Continue using the data from Question 1. Car prices are not only based on manufacture years but also engine size. The engine size (in Liter) for each corresponding sample are given as the following:

Engine size (L)	3.0	(NA)	2.5	(NA)	2.0	3.0	2.5	2.0	3.0	2.0	2.0	(NA)	2.0	3.0	2.0

1. What imputation methods are available in this case? How do you impute the missing values?
2. Use correlation analysis to determine whether **'engine size'** is a useful feature for predicting **'asking price'**. (Hint: Be mindful about the imputation you have used. Think about how imputation may affect the correlation.)

3. Can you suggest any additional preprocessing method? (Hint: What do you think `'NA'` mean?)

4. If you deicide to use both **'year of manufacture'** and **'engine size'** to predict **'asking price'**, what are the preprocessing methods you should apply before building the regression model?

5. We have 3 regression models so far. 2 from Question 1 and we have just built another one. How can you evaluate the performance between these models, and which one do you think has the best performance? (Hint: What Cross-Validation strategy should you apply on a small dataset?)