COMPSCI 762 2022 S1 Week 6 Solution

Luke Chang

April 8, 2022

Question 1

- Apply binning with mean value to split the data into 3 bins based on "Year":

Bin	Mean (Year, Price)	SD (Year, Price)	2SD (Year, Price)
1	1994, 11520	8.0, 6761.1	16.0, 13522.2
2	2008, 15100	1.2, 1546.0	2.4, 3092.0
3	2015, 20180	2.1, 7371.4	4.2, 14742.8

- Absolute error:

Year	9.0	4.0	3.0	7.0	10.0	1.0	1.0	0.0	0.0	2.0	2.0	1.0	0.0	2.0	3.0
Price	280.0	4920.0	10280.0	7120.0	1480.0	500.0	400.0	500.0	1900.0	2300.0	980.0	11980.0	1320.0	7120.0	4520.0

- Based on 95% confidence interval (2σ), there is no outlier in the data.
- The car with lowest price ($4400) seems way cheaper than others. Why isn't it an outlier?
  Note:
  - There are multiple ways determine an outlier. We have to based on a statistical test, not an arbitrary guess. If the criteria we have defined is 95% confidence interval, then none of the data point is outlier.
  - The 1st bin contains older vehicles, and we observe it fluctuates more than other bins.
- If we model the data using least square method ($y = ax + b$), linear transformation will not alter the value of gradient (a). It only changes the bias (b).
- $f(x) = 313.1x - 612397.6$, thus $f(2022) \approx 20691$;
- **Note:** We have to consider *significant figures* (SF) and *decimal point* (DP) when writing the report. The results should keep the least SP. Any parameter we use to compute the result need to keep one additional DP.

Question 2

- Available imputation methods: Median, mode.
- "Engine size" is discrete value. We shouldn't use mean.
- KNN imputation is not suitable in this case, unless we can prove "Year" is correlated to "Engine size".
- If the target of the regression model is "Price", we cannot use "Price" in imputation, because we will not have the target value at inference time.
- Median and mode both give us 2.0.

- Spearman rank correlation in `pandas.DataFrame.corr` can be used.
- Mean price for all 15600; Mean value exclude NA 17900; Mean value for NA only 6400. However, the sample size is too small to consider SD.
- Suggestion: Adding additional feature of "Unknown engine size".
- Ground truth: Engine size is randomly generated, but unknown engine size indicates it is a wrecked car.
- Blindly including extra features may not improve the performance of the model.