Assignment-02

Nuthan reddy vaddi reddy

UTA ID-1002175133

TASK-1:Data from our Lives which a regression model would be appropriate. List some (up to 5) predictors that you might use.

Employee's salary can depend on multiple factors, and a regression model can help understand the relationship between these predictors and an employee's salary.

This are the five predictors that are useful in a regression model to predict employee's salary

Predictor 1: Years of experience.
It is one of the important factor for predicting the employee salary, How many years of real-time experience that the person have will leads to high salary, as experience person will have more real-time experience and skill on the work.

Predictor 2: level of Education.
High level of education degrees may lead to higher salaries like the ph.d student or can get high salary than a undergraduation student, By considering both the level of education and the field of study.

Predictor 3: Location.
The cost of living and taxes varies by different locations. Salaries in areas like california will tend to increase higher for increased living expenses(cost of living) and high taxes. so, person in california will get high salary than person in Texas, so the location is important predictor in the employee salary.

Predictor 4: Job role or position.
Different roles will have different salary within the same organization.For example, Data scienist will earns more than a data analyst.

Predictor 5: Industry or sector of employment.

The industry or sector in which the employee works can have impact on salary. For example, salaries in the technology sector might be higher than those in finance.

TASK-2.1 Munging:

- By using the info() we get the datatypes and number of non-null count of each column.
- In the above data we can see there are 5float64,5int64 and 8object datatypes out of 18 columns.
- In the data body,engine_type,cylinders,bore,stroke,horse_power and peak_rpm columns has numerical data but the datatype is shown as object.
- In the data there is no null values in each column.

- Replace the “?” values with none values by using df.replace in the original dataframe.
- converting the columns bore,store,horse_power and peak_rpm datatypes into float datatypes by using the astype(float) function to analzy the data in the further steps.

- After replacing "?" with None values and converting the columns bore,store,horse_power and peak_rpm datatypes into float datatypes.
- In the above data we can see there are 9float64,5int64 and 4object datatypes out of 18 columns.
- The columns fuel_type,body,wheel_base,length,width,heights,curb_weight,
- engine_type,Cylinders,engine_size,comprassion,city_mpg,highway_mpg and price have no null values.
- The columns horse_power and peak_rpm have 2 null values.
- The columns bore and stroke have 4 null values.

- Droping the columns the body,engine_type and cylinders from the data by using df.drop() and loading the new data into new dataframe df2.

- Dropping the all none values from the original dataframe by using the dropna(inplace=True).
- Where the inplace=True is used for dropping the null values rows from the original dataframe.

- By using isnull().sum() we get count of null values in each column, In the above data the sum of null values in each column is zero.
- converting the categorical variables into binary (0 or 1) columns.
- columns=['fuel_type'] This specifies the column in the DataFrame df2 in the original dataframe.
- After completing all the steps in task 2.1, the data we can see there are 9float64,5int64 and 1uint8 datatypes out of 15 columns.
- All the columns in the data will have no null values.

TASK-2.2 EDA on df2:

Exploratory Data Analysis Report(EDA)

1.Provide descriptions of your sample and features:

- Finding the shape of the data by using df2.shape, In the above data shape is (195,15) 195-rows and 15-columns
- In the data we can see there 195-rows and 15-columns, In which there are 9-float64,5-int64 and 1-unit8 datatypes.
- The number of non-null values in each column is zero.
- df2.describe() command used to find mean count standard deviation,min,max,count and quartlies of the data.

2. Check for missing data:

- In this step we checking the null values , duplicate values in the dataset and removing the null values and duplicate value from the data.
- By using isnull().sum() we get the sum all null in each column, In the data the sum of all null values in each column is zero.
- By using df2.drop_duplicates() we drop the duplicate values from the dataframe, In the data 3 rows are dropped from the dataframe.
- By using df2.fuel_type.unique(), we get only the unique values gas and diesel from the fuel_type column.

- By using df2.value_counts("fuel_type_gas") we get the count of values in the fuel_type column
- 1-175 and 0-20
- In the above data we can see the dataset value count of each class is not same, so the dataset is not balanced and it varies excessively and it is not good for prediction.
- For prediction the same value count in each class is very important.
- The accuracy of prediction is high when we have same value count in each class.
- For prediction the same value count in each class is very important.
- The accuracy of prediction is high when we have same value count in each class.

3. Identify the shape of your data:

- These observations provide a preliminary understanding of the central tendencies, variability, and ranges of the numerical features in the dataset.
- By sns.countplot(x="fuel_type_gas", data=df2,) we get the count plot for the values in the fuel_type_gas column.
- 1-175 values means gas
- 0-20 values means diesel

- we can see the dataset value count of each class is not same, so the dataset is not balanced.
- The stripplot is used specify that wheel_base is on the y-axis and fuel-type is on the x-axis.
- In The above stripplot we can see observe that the gas(1) has the highest wheel_base with above 120 and the diesel(1) has the low wheel_base lies above 115 .
- gas(1) data mostly lies between 90 to 105 wheel_base.

- The violinplot is used specify that length is on the y-axis and fuel-type is on the x-axis.
- In The above stripplot we can see observe that the gas(1) has the highest length with 210 and the diesel(0) has the low length lies above 200 .

- gas(1) data mostly lies between 160 to 190 length.

4. Identify significant correlation:

- Correlation Analysis: I have extracted the heatmap of my data set to find the correlation between the columns of my datset.

- wheel_base and length,width,heights,curb_weight, Engine_size,bore,stroke,comprassion,horse_power, price has positive Correlation.
- wheel_base and peak_rpm,city_mpg,highway_mpg has negative correlation.

- length and wheel_base,width,heights,curb_weight,engine_size, bore,stroke,comprassion,horse_power,price has positive Correlation.
- length and peak_rpm,city_mpg,highway_mpg has negative correlation.

- width and wheel_base,length,heights,curb_weight, engine_size,bore,stroke,comprassion,horse_power,price has positive Correlation.
- width and peak_rpm,city_mpg,highway_mpg has negative correlation.

- Height and stroke,peak_rpm,city_mpg,highway_mpg,horse_power has negative correlation.
- Height and length,width,heights,curb_weight,engine_size,bore, comprassion,price has positive Correlation.

- curb_weight and length,width,heights,wheel_base,engine_size, bore,stroke,comprassion,horse_power,price has positive Correlation.
- curb_weight and peak_rpm,city_mpg,highway_mpg has negative correlation.

- engine_size and length,width,heights,curb_weight, wheel_base,bore,stroke,comprassion,horse_power,price has positive Correlation.
- engine_size and peak_rpm,city_mpg,highway_mpg has negative correlation.

- bore and length,width,heights,curb_weight,wheel_base,engine_size, comprassion,horse_power,price has positive Correlation.
- bore and peak_rpm,city_mpg,highway_mpg,stroke has negative correlation.

- stroke and length,width,curb_weight,wheel_base,engine_size, comprassion,horse_power,price has positive Correlation.
- stroke and peak_rpm,city_mpg,highway_mpg,bore,heights has negative correlation.

- comprassion and length,width,curb_weight,wheel_base,engine_size, stroke,city_mpg,highway_mpg,bore,heights,price has positive Correlation.
  Comprassion and peak_rpm,horse_power has negative correlation.

- horse_power and length,width,curb_weight,wheel_base, engine_size,stroke,bore,price has positive Correlation.
- horse_power and city_mpg,highway_mpg,comprassion,heights has negative correlation.

- peak_rpm and horse_power has positive Correlation.
- peak_rpm has negative Correlation with all columns expect horse_power.

- city_mpg and comprassion,highway_mpg has positive Correlation.
- city_mpg has negative Correlation with all columns expect comprassion,highway_mpg.

- highway_mpg and city_mpg has positive Correlation.
- highway_mpg has negative Correlation with all columns expect city_mpg.

- price and length,width,heights,curb_weight,engine_size,bore,stroke, comprassion,horse_power,wheel_base has positive Correlation.
- Price and peak_rpm,city_mpg,highway_mpg has negative correlation.

5.Spot/deal with outliers in the dataset:

- I have extracted the boxplot for the all columns and the columns which are having the outliers are removed from the data because The outliers will affect model accuracy ,effect results and not good for data analysis.

- By Box plot for wheel_base we can find the two outliers which differ from the majority of the data and has extreme values. The outliers will affect model accuracy and effect results, so removing outliers from the data.

- By Box plot for width we can find the three outliers which differ from the majority of the data and has extreme values.
- The outliers will affect model accuracy and effect results, so removing outliers from the data.

- By Box plot for city_mpg we can find the two outliers which differ from the majority of the data and has extreme values.
- The outliers will affect model accuracy and effect results, so removing outliers from the data.

- By Box plot for highway_mpg we can find the two outliers which differ from the majority of the data and has extreme values.
- The outliers will affect model accuracy and effect results, so removing outliers from the data.

- By Box plot for engine_size we can find the five outliers which differ from the majority of the data and has extreme values.
- The outliers will affect model accuracy and effect results, so removing outliers from the data.

- By Box plot for horse_power we can find the three outliers which differ from the majority of the data and has extreme values.

- The outliers will affect model accuracy and effect results, so removing outliers from the data.

- By Box plot for stroke we can find the six outliers which differ from the majority of the data and has 4 extreme values and 2 low values.
- The outliers will affect model accuracy and effect results, so removing outliers from the data.

- By Box plot for comprassion we can find the nine outliers which differ from the majority of the data and 8 are extreme values and one is low values.
- The outliers will affect model accuracy and effect results, so removing outliers from the data.

- By Box plot for peak_rpm we can find the one outliers which differ from the majority of the data and extreme value.
- The outliers will affect model accuracy and effect results, so removing outliers from the data.

- By Box plot for price we can find the 12 outliers which differ from the majority of the data and extreme values.
- The outliers will affect model accuracy and effect results, so removing outliers from the data.

- By Box plot for heights we can find there no outliers in the data.The heights data is good for data analyzing.

- By Box plot for curb_weight we can find there no outliers in the data.The curb_weight data is good for data analyzing.

- By Box plot for bore we can find there no outliers in the data.The bore data is good for data analyzing.

After, Explority data analysis the data have no null values,no missing values, no outliers and no duplicates in the dataset. The dataset is clean and ready to use for the data analysis and future predictions.

Task 3 multiple Regression Analysis:

3.1-Create a model that uses all the variables

1-The intercept in the regression model signifies the estimated dependent variable (price) when all independent variables are at zero. However, in the data, variables like engine size, height, and width are never zero. In our model, the intercept is -6.207e+04

2-P>|t|" (p-value) column p-values below a significance level is considerd as 0.05 are considered statistically significant are width,heights,engine size,stroke,comprassion,horse_power,peak_rpm.

3-Model variance assesses the disparity between actual values and predicted values of the dependent variable (price). The model's R-squared value, such as 0.857, the proportion of price variance elucidated by independent variables. An R-squared of 0.857 implies that roughly 85.7% of the price variability is accounted for by the model.

4- determination:The coefficient of determination, often called the R-squared, is a measure of how well the independent variables explain the variation in the dependent variable. In this model, R squared is 0.857. This means that about 85.7% of the variation in car prices is explained by the independent variables in the model. A higher R-squared indicates a better fit of the model to the data. However, keep in mind that a high R-squared does not necessarily mean that the model's predictions are always accuratecoefficeint of.

5-The coefficient of determination, R-squared, quantifies the extent to which independent variables clarify the variance in the dependent variable. Here, R-squared is 0.857, signifying that approximately 85.7% of car price variation is accounted for. A higher R-squared implies a superior model fit, but it doesn't guarantee perfect predictions.

3.2-Drop all the variables that are not statistically significant:

1-The intercept (const) signifies the estimated price when all independent variables are at zero. However, in this context, it implies an unrealistic scenario where multiple variables are set to zero, leading to a negative estimated price of around -$64,000, which may not align with the dataset's context.

2- A variable is seen as important in the analysis when its p-value is significance level is considerd as 0.05 the values below 0.05. In your findings, all independent variables, except "altitude," are important because their p-values are less than 0.05.
3-With an R-squared of 0.853, approximately 85.3% of price variation is accounted for by the model's independent variables, indicating its effectiveness in explaining price fluctuations. However, R-squared doesn't assess prediction quality or potential overfitting, leaving those aspects unaddressed.
4-The adjusted R-squared (0.847) is a variant of R-squared, accounting for model complexity by considering independent variables. It discourages overfitting by variables. Its slight reduction compared to R-squared, common with numerous variables, offers a more cautious assessment of the model's data fit.
5-The F-statistic tests the model's overall significance. With an F-statistic of 154.5 and a low p-value (3.25e-74), the model is highly significant. This implies that at least one independent variable significantly influences the dependent variable. The high R-squared and low p-value indicate a strong fit, but model appropriateness depends on specific research context and regression assumptions.

3.3 Compare the two models with ANOVA:

1-Null hypothesis (H0): Model 1 is as effective as, or superior to, Model 2 in elucidating the variability of the dependent variable.
2-Alternative Hypothesis (Ha): Model 2 exhibits a significant enhancement in explaining the variance of the dependent variable in comparison to Model 1 (the reduced model).
To assess the hypothesis, individuals often scrutinize the p-value (Pr(>F)) associated with the F statistic presented in the ANOVA table. If the p-value falls below the predetermined significance level, typically set at 0.05, the null hypothesis is discarded. In simpler terms, if Pr(>F) is less than 0.05, this leads to the conclusion that a substantial distinction exists between the two models, affirming that the full model (model 1) offers a more accurate fit. so,since an exact p-value for your ANOVA outcome has not been provided, it is essential to compare it with our chosen significance level to make a conclusive determination.

3.4 Checking the assumptions:

1. Linearity of Relationships:
Assumption: The relationship between independent and dependent variables is linear.
hold: This holds true if scatterplots show nearly linear trends for each independent variable.

2. Independence of Residuals:
Assumption: Residuals (observed - predicted values) are independent.
hold : Assessing residual independence is essential, although not explicitly tested.

3. Homoscedasticity:
Assumption: Residual variance is constant across independent variable levels.
hold: This is met if residuals vs. predicted values exhibit a relatively uniform spread.

4. Normality of Residuals:
Assumption: Residuals follow a normal distribution.
hold : For large samples, deviations from normality are tolerable. Check normality using histograms or Q-Q plots.

5. Multicollinearity:
Assumption: Independent variables are not highly correlated.
hold: Evaluate using correlation coefficients or variance inflation factors (VIF).

6. Exogeneity:
Assumption: Independent variables are uncorrelated with residuals.
hold: Address violations by including relevant variables in the model as needed.

3.5 Is there Multicollinearity in your data:

The VIF analysis reveals that "width," "height," "stroke," "horse_power," and "peak_rpm" exhibit high VIF values, well beyond 10. "engine_size" demonstrates moderate multicollinearity with a VIF of approximately 65. In contrast, "compression" has a relatively lower VIF of around 10, indicating less multicollinearity in this variable.

The Variance Inflation Factor (VIF) is a metric for assessing multicollinearity among independent variables in a regression model. Multicollinearity arises when these variables are highly correlated, hindering the isolation of their individual impact on the dependent variable. VIF values exceeding 10 typically signal troublesome multicollinearity.