**NikhilKumar Patel**

**Assignment-based Subjective Questions**

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

1. Why is it important to use **drop\_first=True** during dummy variable creation? (2 mark) [Answer]: Pandas has an inbuilt method to create the dummy variables, where we specify drop\_first=True telling pandas that it has to remove the variable respecting the postulate for categorical variables with n-levels, it needs n-1 variables to represent the dummy variables. This will help in reducing the correlation created among dummy variables.
2. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

[Answer]: Temp & Registered

1. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

[Answer]:

Linear relationship exists between the dependent variable and the predictors. This was verified at first by using the scatterplot and pairplot.

Checked if the errors are normally distributed or not.

Checked the p-value and VIF values with the following interpreataions:

\* The p-value is 0, therefore, it is a good sign.

\* R2 value is 28% which means that when take into account area as one of the variable - there is 28% variance in the price of the house

1. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

[Answer]: Based on the model, Temperature, Summer & Winter, September month are contributing significantly towards the demand of the shared bikes.

**General Subjective Questions**

1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Chart, scatter chart

Description automatically generated

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

**Y = beta0 + beta1 \* x**

While training the model we are given :

x: input training data (univariate – one input variable(parameter))

y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x.

The model gets the best regression fit line by finding the best beta1 and beta2 values. **beta1: intercept**

**beta2: coefficient of x**

Once we find the best beta1 and beta2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

How to update beta1 and beta2 values to get the best fit line ?

Cost Function (J): By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the beta1 and beta2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).

Gradient Descent:

To update beta1 and beta2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random beta1 and beta2 values and then iteratively updating the values, reaching minimum cost.

1. Explain the Anscombe’s quartet in detail.

{Answer]:

Chart, scatter chart

Description automatically generated

Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

1. What is Pearson’s R?

[Answer]:

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

1. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

[Answer]:

Scaling is a subjective discussion on making the numbers feel right, e.g. between zero and one, or one and a hundred. For example converting data given in millimeters to meters because it's more convenient, or imperial to metric.

While normalisation is about scaling to an external 'standard' - the local norm - such as removing the mean value and dividing by the sample standard deviation, e.g. so that your sorted data can be compared with a cummulative normal, or a cummulative Poisson, or whatever.

1. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

[Answer]:

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

1. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

[Answer]:

Q Q Plots (Quantile-Quantile plots) are plots of two [quantiles](https://www.statisticshowto.com/quantile-definition-find-easy-steps/)against each other. A quantile is a fraction where certain values fall below that quantile. For example, the [median](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median)is a quantile where 50% of the data fall below that point and 50% lie above it.

**Why we use Q-Q Plot?**

The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

Chart, line chart

Description automatically generated

The image above shows quantiles from a theoretical [normal distribution](https://www.statisticshowto.com/probability-and-statistics/normal-distributions/) on the horizontal axis. It’s being compared to a set of data on the y-axis. This particular type of Q Q plot is called a normal quantile-quantile (QQ) plot. The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.