Assignment 9

1. Explain One-Hot Encoding and Label Encoding. Does the dimensionality of the data set increase or decrease after encoding, if yes then how?

Using one-hot encoding increases the dimensionality of the data set having categorical features. Label encoding doesn't affect the dimensionality of the data set. One-hot encoding creates a new variable for each level in the variable whereas, in Label encoding, the levels of a variable get encoded as 1 and 0.

2. What is Target Encoding and how it is different from one hot encoding?

Target encoding is the process of replacing a categorical value with the mean of the target variable. Any non-categorical columns are automatically dropped by the target encoder model. You can also use target encoding to convert categorical columns to numeric.

Example for target encoding:

	Animal	Target	Encoded Animal
0	cat	1	0.40
1	hamster	0	0.50
2	cat	0	0.40
3	cat	1	0.40
4	dog	1	0.67
5	hamster	1	0.50
6	cat	0	0.40
7	dog	1	0.67
8	cat	0	0.40
9	dog	0	0.67

Calculating the probability of occurrence of a particular class

	Animal Group	Target 0	Target 1	Probability of 1
0	cat	3	2	0.40
1	dog	1	2	0.67
2	hamster	1	1	0.50

Example of One-hot Encoding

This type of encoding simply produces one feature per category, each binary. Hence increases the number of columns. If a particular class exits it is assigned 1 and rest 0.

	Animal	Target	Animal Encoded	isCat	isDog	isHamster
0	cat	1	0	1.0	0.0	0.0
1	hamster	0	2	0.0	0.0	1.0
2	cat	0	0	1.0	0.0	0.0
3	cat	1	0	1.0	0.0	0.0
4	dog	1	1	0.0	1.0	0.0
5	hamster	1	2	0.0	0.0	1.0
6	cat	0	0	1.0	0.0	0.0
7	dog	1	1	0.0	1.0	0.0
8	cat	0	0	1.0	0.0	0.0
9	dog	0	1	0.0	1.0	0.0

3. If you have a date column in our dataset, then how will you perform Feature Engineering in pandas?

If we have a date a column:

- Check for the data type is object or Date Time.
- If it is object type, it is then converted to Date Time format. Pandas has a built-in function called to_datetime()that converts date and time in string format to a DateTime object.
- We can perform a time series analysis by setting the index column to Date. This can be done using set_index() function. df= df.set_index('Date') ## if df is the data set.
- Generate sequences of fixed-frequency dates and time spans.
- Manipulating and converting date times with timezone information.
- Resampling or converting a time series to a particular frequency.
- Finding the particular day form a date using Timestamp()

4. How do you perform feature selection with Categorical Data?

Feature selection is the process of identifying and selecting a subset of input features that are most relevant to the target variable.

There are two popular feature selection techniques that can be used for categorical input data and a categorical (class) target variable.

They are:

Chi-Squared Statistic: The results of chi-square test can be used for feature selection, where those features that are independent of the target variable can be removed from the dataset.

Mutual Information Statistic: Mutual information is the application of information gain (typically used in the construction of decision trees) to feature selection. Mutual information is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable.

5. When would you remove Correlated Variables?

In a more general situation, when we have two independent variables that are very highly correlated, we must remove one of them because we run into the **multi-collinearity** conundrum and our regression model's regression coefficients related to the two highly correlated variables will be unreliable.
To remove the correlated features, we can make use of the corr() method of the pandas dataframe. The corr() method returns a correlation matrix containing correlation between all the columns of the dataframe. And remove any one from the correlated pairs.