CMPE 255 – Data Mining- assignment #1

The operations performed in the Colab notebook is described here.

* Firstly, the initially required libraries are imported to read the data set and to perform basic data cleaning operations. The libraries used are pandas, numpy, matplotlib.
* The dataset is available as a csv(comma separated values) file. The dataset can be found at <https://www.kaggle.com/jessemostipak/hotel-booking-demand>**.**
* The dataset describes the booking data from a hotel booking website. The dataset has features like the type of hotel, number of people staying, lead time which describes how far ahead that the reservation was made. Importantly, it includes the cost of each room. This notebook tries to predict the price per day per person based on parameters like leadtime, room type, meals and other special requests.
* The dataset is read into the colab notebook using pandas. Initially it has 119390 rows and 32 columns.
* The dataset is noisy. It must be cleaned. We start of by finding null values. There are null values in few of the features. We are also looking for unique values. Please refer to the dataset description file which describes all the columns. Some columns are categorical data. We are using the unique value function to make sure that the unique values are the same as the categories.
* We are dropping few features which are not necessary for our model. This also takes care of some of the null values.
* We are also replacing null values with 0 for other numerical features.
* For categorical data, we are comparing the unique values with the pre-existing list of categories. Some values are generalized. Eg : Online agent and offline agent are combined as agent booking.
* We are creating new features for total nights spent in the hotel using two features namely, no of weeknights spent and number of weekend nights spent.
* Another feature called price per person per day created using the total nights feature from the previous step.
* Null and inf values are handled again.
* We are testing for outliers in the dataset. Outliers are eliminated using the mean and standard deviation method. Any value outside the 1 standard deviation from the mean is removed.
* Using a scatter plot of the arrival week of the year and the cost per day scatter plot we can find outliers. Two scatter plots visualize the dataset before and after the outlier removal process.
* The filtered dataset includes categorical data. Onehot encoding is used to encode categorical data with less categories. Binary encoding is used to encode features with a large number of categories.
* The dataset is now split into training and testing data set with a ratio of 0.2 for training and testing.
* Firstly, we use a linear regressor model to predict the cost per person per day. We achieve an accuracy score of 42.3%
* The gridsearch method is used to compare different models. Linear regression, lasso and decision tree algorithms are implemented in the gridsearch method.
* The project has achieved a highest accuracy of 83.9% using decision tree regressor.

Please find the notebook on GitHub : <https://github.com/sharadnatarajSJSU/CMPE255Assignment_1>

The repository is fully accessible to public.