CS 663 - Machine Learning Spring, 2021

Lab 01 – Property Permits - Pandas

This dataset provided comprises of property permits across San Francisco. The fields include details on application/permit numbers, job addresses, supervisorial districts, and the current status of the applications.

In this lab, you are expected to perform an EDA on the given dataset, do some pre-processing and identify fields which have linear correlation. Using the linear relation, helps to understand the combination of fields (2 or more) which are optimal for modelling. You are expected to use techniques in pandas to manipulate the dataset, visualize it using Matplotlib/Seaborn/Plotly and reach a conclusion.

Extra credit (5%) - There are two optional tasks listed. The assignment is capped at **100 pts**. You can use this to make up for lost points from other parts.

Data

The link dataset:
Name: **property_permits.csv**

Process

Your implementation must:

- Browse to (github.classroom) and start a repo from there. (In the future, there may be "starter code" for you.) Choose your name and accept the assignment. If you do not see your name, contact the instructor / TA.
- Use a Jupyter notebook or Python script to complete your implementation.
- Load the dataset, which you may assume is in the same directory as the notebook / script.
- **EDA** - Inspect the dataset to understand the different attributes and their use case. Remove fields that have majority of records as (null)s. (80-90%)
- [**Optional**] As part of **pre-processing**, you can choose to deal with (null) values using some generalization. An example would be to distribute values based on the StreetNames or DistrictNames in San Francisco. You may find a better way to do this.
- Select a use case of choice and plot a histogram for the same. One example could be distribution of distinct permit types across the dataset. Normalize the ranges for clearer understanding.
- Plot **visualizations** representing relationship between potential attributes (combination of two fields) that can be utilized for linear modelling and predictions. You can make use of box / scatter / line plots.
- Once the linear relation between any two fields have been identified, use a **correlation matrix** or heat map to validate the same. State which columns would you like to select as inputs and target for modelling. (df.corr(), assuming df is the dataframe object. You can convert it into a heat map using seaborn for clearer visualization)
- [**Optional**] Identify a set of attributes (3 or more), that can be used for modelling. Specify the inputs to the model (2 or more) and target field to be predicted.
- There isn't an expected approach for implementing the above. It will be interesting to understand your individual perspective while working on the given problem.

Grading

Grades for this assignment will be determined by the grader as follows:

- 100% = Code functions, is well-documented and clearly shows the relationship between fares and survival.
- 75% = Code functions but is not well-documented or does not clearly show the relationship between fares and survival.
- 50% = Code functions but is not well-documented -AND- does not clearly show the relationship between fares and survival.
- 0% = No submission / code does not function.

Submission

Submit the link to your GitHub repository on Canvas.