Applied Competitive Lab in Data Science

Exercise 2

Deadline: 10.11.2022 at 23:55

0 Submission Guidelines

Your grade would be made from the top 3 questions you've answered.

Submit a single Jupyter notebook file containing your answers(including your plots, code, and textual answers). Your notebook should be run from beginning to end without errors, and conform to the style shown in class.

For questions 1-3 use the clean version of exercise 1's dataset(ex2.csv). For question 4 use the covid19 dataset. Both are available in the exercise page.

1 Question 1

(a) In exercise 1, we performed a rather naive imputation for null values using the mode statistic. This may be fine for some discrete variables, but not always so for continuous variables. Now, attempt to use what we've seen in class to find a better imputation strategy for one continuous and one categorical feature. What did you choose? And why?

(b) In class we've talked about the process of binning features as a way to avoid having to account for very rare categories and as a method of creating new, more intuitive overall features. Select one of the dataset's features, like any high-cardinality categorical feature (i.e any categorical feature with many different categories) or continuous feature you could summarize, as well as binning strategy, and implement it. Explain you choice.

2 Question 2

(a) At times, we'd like to be able to expand the information that exists in our dataset. For this question, try to think of additional information that could prove useful when analyzing the data in the given dataset. What sort of information would you add?

(b) Use any online dataset you might think suitable and add a new feature to the dataset using it. You may use any source you'd like, but a few examples and suggestions are the datasets found below. Both have columns to use to merge them to your own data, as well as a description of each feature:

   (a) https://www.kaggle.com/benhamner/2016-us-election - Which contains census data relating to the 2016 elections

   (b) https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations - Which contains data relating to household income

3 Question 3

(a) In class, we've seen the use of aggregative functions and features as an option for feature addition. Which aggregations would you add to the dataset? Suggest and add two aggregative features (make sure you aggregate over more than a single value overall- don't calculate all your suggested features over the same column). You may also use any other feature you suggested above or calculated yourself as the base for the aggregation. Explain why you decided to add these features.

(b) Show the correlation heatmap for your overall dataset (if you've added new features in previous subquestions- keep them)- are there any new surprising correlations?

4 Question 4

(a) Try to fit a linear regression model using the day feature to predict the amount of new detected cases. You don't need to perform any train test split for this. Report the mean squared error metric for your model- you may implement the linear regression model and error metric calculation yourself, but you may (and should) also use the existing implementations from sklearn.

(b) Now, visualize the connection between the independent variable and the dependent variable. What sort of relationship exists between them? Think of how this relationship would hold in the long run- not just for the data we've seen.

(c) Assess the relationship you mentioned in the previous question- visualize it, and according to this relationship try to fit a new model. You may attempt to use several models using several different transformations- try and find the one that gives you the best results. Report the result.