UDACITY

‹ Return to Classroom

Machine Learning Capstone Project

REVIEW

HISTORY

Meets Specifications

Dear student,

This is a very cool analysis and a great read. You have demonstrated a full understanding of the entire machine learning pipeline. You really took the time to go above and beyond in all the requirements.

Hopefully, you have learned a bunch throughout this capstone project and you can take some of these techniques further.

I wish you all the best.

Keep it up!

Definition

Student provides a high-level overview of the project in layman's terms. Background information such as the problem domain, the project origin, and related data sets or input data is given.

Great work here. Congrats

The problem which needs to be solved is clearly defined. A strategy for solving the problem, including discussion of the expected solution, has been made.

You have defined clearly the problem statement here. Congrats!

Metrics used to measure performance of a model or result are clearly defined. Metrics are justified based on the characteristics of the problem.

Excellent choice of metrics and good explanation. Congrats!

Suggestions and comments

I think these links might be useful:

- https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce
- https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Analysis

If a dataset is present, features and calculated statistics relevant to the problem have been reported and discussed, along with a sampling of the data. In lieu of a dataset, a thorough description of the input space or input data has been made. Abnormalities or characteristics about the data or input that need to be addressed have been identified.

Very nice job describing your dataset.

Suggestions

Maybe also look into computing the Kolmogorov-Smirnov test for goodness of fit. (https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.kstest.html)

A visualization has been provided that summarizes or extracts a relevant characteristic or feature about the dataset or input data with thorough discussion. Visual cues are clearly defined.

All great visuals to show here! The distribution of features and correlation plots are always a good idea. Maybe another idea would be to combine the target variable with these features to get a better understanding of how each correlates to the target variable with a seaborn factorplot.

Algorithms and techniques used in the project are thoroughly discussed and properly justified based on the characteristics of the problem.

Student clearly defines a benchmark result or threshold for comparing performances of solutions

obtained.

Methodology

All preprocessing steps have been clearly documented. Abnormalities or characteristics about the data or input that needed to be addressed have been corrected. If no data preprocessing is necessary, it has been clearly justified.

Excellent job documenting all your pre-processing steps. Great idea to use.

Suggestions

Could also check out the 'normality' of the features with a quantile-quantile (q-q) plot (https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm) in terms of why a log transformation would be ideal.

```
import scipy
import matplotlib.pyplot as plt
for feature in data.columns:
    scipy.stats.probplot(data[feature], plot=plt)
    plt.title(feature)
    plt.show()
```

The process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented. Complications that occurred during the coding process are discussed.

Very solid step by step process here, as it is quite clear in how you approached this problem. Your results would definitely be replicable.

The process of improving upon the algorithms and techniques used is clearly documented. Both the initial and final solutions are reported, along with intermediate solutions, if necessary.

Results

The final model's qualities — such as parameters — are evaluated in detail. Some type of analysis is used to validate the robustness of the model's solution.

Suggestions

Suggestions

Maybe one other idea would be to plot and 95% confidence interval to determine if the model is robust with bootstrapping. Here might be an example with the boston housing dataset.

```
from sklearn.utils import resample
import matplotlib.pyplot as plt
data = pd.read_csv('housing.csv')
values = data.values
# configure bootstrap
n_iterations = 1000
n_size = int(len(data) * 0.50)
# run bootstrap
stats = []
for i in range(n_iterations):
    # prepare train and test sets
    train = resample(values, n_samples=n_size)
    test = np.array([x for x in values if x.tolist() not in train.tolist()])
    # model
    model = DecisionTreeRegressor(random_state=100)
    model.fit(train[:,:-1], train[:,-1])
    score = performance_metric(test[:,-1], model.predict(test[:,:-1]))
    stats.append(score)
# confidence intervals
alpha = 0.95
p = ((1.0 - alpha) / 2.0) * 100
lower = max(0.0, np.percentile(stats, p))
p = (alpha + ((1.0 - alpha) / 2.0)) * 100
upper = min(1.0, np.percentile(stats, p))
# plot
plt.hist(stats)
plt.axvline(lower, color='red', lw=3)
plt.text(lower, n_iterations // 4, 'Lower Bound', rotation=90)
plt.axvline(upper, color='red', lw=3)
plt.text(upper, n_iterations // 4, 'Upper Bound', rotation=90)
plt.title('Distribution of Classification Accuracy Using the Bootstrap')
plt.show()
print('%.1f confidence interval %.1f%% and %.1f%%' % (alpha*100, lower*100, u
pper*100))
```

The final results are compared to the benchmark result or threshold with some type of statistical analysis. Justification is made as to whether the final model and solution is significant enough to have adequately solved the problem.

Conclusion

A visualization has been provided that emphasizes an important quality about the project with thorough discussion. Visual cues are clearly defined.

Excellent job!

Student adequately summarizes the end-to-end problem solution and discusses one or two particular aspects of the project they found interesting or difficult.

Discussion is made as to how one aspect of the implementation could be improved. Potential solutions resulting from these improvements are considered and compared/contrasted to the current solution.

Quality

Project report follows a well-organized structure and would be readily understood by its intended audience. Each section is written in a clear, concise and specific manner. Few grammatical and spelling mistakes are present. All resources used to complete the project are cited and referenced.

Great job here.

Code is formatted neatly with comments that effectively explain complex implementations. Output produces similar results and solutions as to those discussed in the project.

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review