UDACITY

PROJECT

Machine Learning Capstone Project

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

13 SPECIFICATIONS REQUIRE CHANGES

This is a very interesting report. One tip here would be to make this report more personable and not just directly about the implementation, as it would be a good idea to use the context of the problem more here.

Thus make sure you look at the comments provided here and the project_report_template. As this would be good to follow. Look forward in seeing your next submission!!

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.

You seem to just dive into the problem statement here. Therefore this opening section should get the reader excited about the project, thus please provide a high-level overview of the project in layman's term.

Also when you mention the data set, please state where it was obtained from? Did this come from kaggle? Online site? Etc...

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.

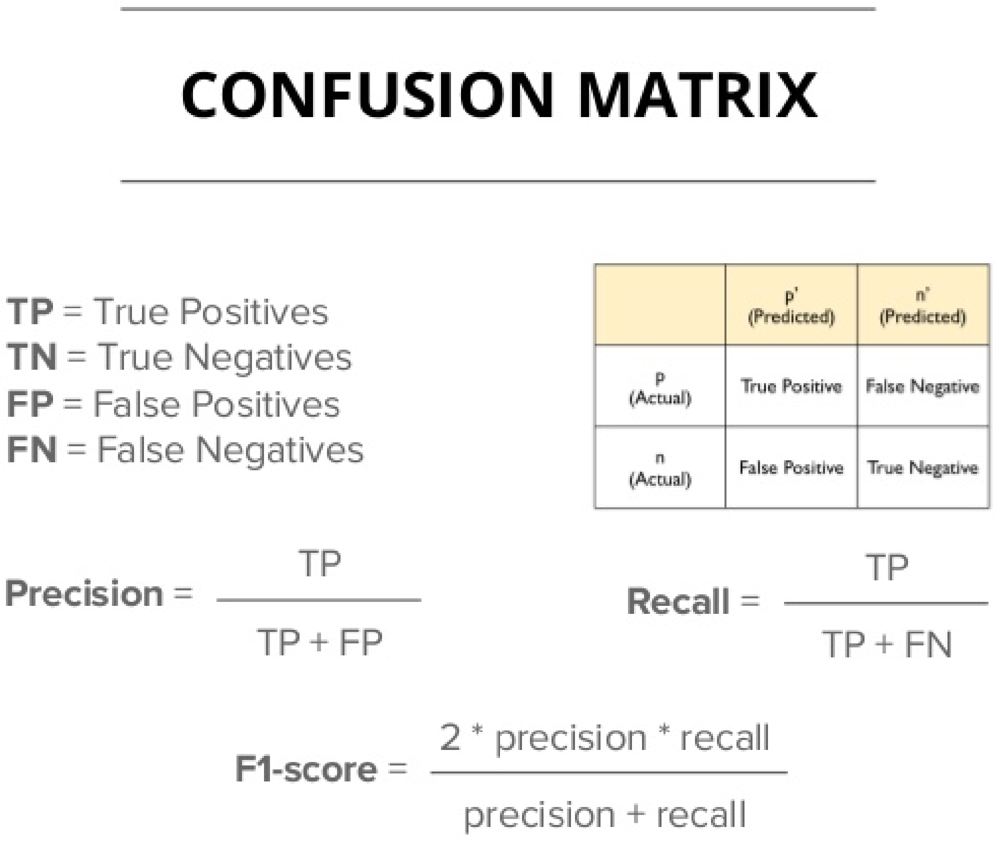
"*The objective is to predict the segment for a new customer using a classifier.*"

Your problem state is well defined and you have given a good strategy for solving the problem and it is clear that you are trying to predict Backpackers, Family, or Couple.

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

Can you expand on "*This F1 score is the metrics for this project.*" Thus why is an F1 score appropriate for this report?

Additional Ideas: Would also be nice to see a thorough description of what exactly a F1 score calculates.



CONFUSION MATRIX

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

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.

Good job describing the columns descriptions, and giving an idea of the number of rows / columns in the individual datasets and the combined. And once you have combined both of the datasets you have provided a sample of the data.

Therefore the last thing you should do here is also give some calculated statistics relevant to the problem. As after

"*Now the X_train consists of the following 10 features,*
1> Age (3), ..."

would be a good place to show and describe a `data.describe()`

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.

The PCA plot and the ICA plots are really cool here and gives a good idea of the correlation and variance of the dataset / independent features.

Interesting that you have decided to apply PCA on this dataset as 10 features can still be handled pretty successfully.

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

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain

Therefore I see you have tried a Decision Tree, Logistic Regression, KNN, AdaBoost, etc.. Therefore why we these considered here. As go model by model and briefly describe how each works and why it would be suitable for such a problem?

Also make sure you discuss your "*KBest, Principal Component ANalysis (PCA) and Independent Component Analysis (ICA) are implemented.*" in more detail here as well.

Note: I would recommend reducing the number of different classifiers.

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

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed.

Therefore what would be a good F1 score benchmark for this problem? Are there are other online F1 scores out there that you can try and beat?

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 with the preprocessing steps needed for this problem, as it seems like this is where the most time has been done. Thus anytime we need to combine multiple dataset, this typically does happen.

As you know what they say a machine learning engineer spends more of their time cleaning data and preprocessing features.

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.

Good idea to state the different functions created and the train / test split. However for this section you should also address any complications that occurred during the coding process. Where there any issues?

Additional Ideas:

- What are data split are you using here? 70/30? Why did you use `num_test = 40000`?
- Some code snippets could also be a nice touch

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.

Why was "*1> Decision Tree Classifier 2> Random Forest Classifier 3> KNeighbors Classifiers*" the "*BEST THREE classifiers for further analysis.*"? What are you evaluating this on? Solely F1 score? training time and F1 scores? Are we overfitting on any? Etc..

Great idea to use GridSearch / RandomGS, please also discuss these results of the fine tuned models. Do they improve? Why? Etc..

Additional Ideas:

- Why use random CV?

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.

"*The best one chosen is,*
DecisionTreeClassifier with the following parameters,
('min_samples_split': 10, 'max_leaf_nodes': 10, 'criterion': 'gini', 'max_depth': 'None', 'min_samples_leaf': 5)"

Therefore please examine these in detail here. Also why was Decision tree chosen over KNN and RF? As it should be clear how the final model was derived and why this model was chosen.

Also for this section, please provide some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model's solution is affected (this is called sensitivity analysis). As

- Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?
- Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?
- Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?
- Can results found from the model be trusted?

Thus please provide some validation for the model.

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.

Once you implement a benchmark. In this section, your model's final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project.

Conclusion

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

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss.

Since you are using a decision tree or random forest a feature importance plot would be a good idea `clf.feature_importances_`

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

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

- Have you thoroughly summarized the entire process you used for this project?
- Were there any interesting aspects of the project?
- Were there any difficult aspects of the project?
- Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?

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.

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

- Are there further improvements that could be made on the algorithms or techniques you used in this project?
- Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?
- If you used your final solution as the new benchmark, do you think an even better solution exists?

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.

Easy to read, I would recommend following the guidelines in the project_report_template

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

Code Note: In your code

```
    grid_search = GridSearchCV(clf,
                               param_grid=param_grid,
                               cv=cv, verbose=10)
```

You are evaluating this based on accuracy(default param) and not F1_score, therefore check out sklearn.metrics.make_scorer and create your own scoring metric and use a F1 score instead

☑ RESUBMIT

DOWNLOAD PROJECT

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Rate this review

Student FAQ