UDACITY

‹ Return to Classroom

Machine Learning Capstone Project

审阅 | 代码审阅 | HISTORY

符合要求

Dear student,
The report is clear, well organized, readable, and easy to understand. Very impressive submission here, as you have good understanding of these techniques and you now have a solid understanding of the machine learning pipeline. Check out corresponding sections based on your submitted comment. Hopefully you can learn a bit more from this review. I have been reading this paper recently and found it very interesting. I hope you will too. All the best with your future endeavours!

Some resources you may want to consider referring to:

- For deep learning - https://www.fast.ai/
- General ML, more Developer Oriented - https://machinelearningmastery.com/
- General Data Science - https://www.datasciencecentral.com/
- Podcast - https://dataskeptic.com/
- Grokking Machine Learning - our Luis Serrano's Intro Level Book https://www.manning.com/books/grokking-machine-learning
- Learning from Data - CalTech Professor abu-Mostafa's lectures http://work.caltech.edu/lectures.html
- Best AI papers 2020: https://github.com/louisfb01/Best_AI_paper_2020
- https://github.com/Developer-Y/cs-video-courses#math-for-computer-scientist
- https://deep-learning-drizzle.github.io/
- https://farid.one/kaggle-solutions/

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.

Nice overview of the problem domain! I love the focus on the real-world impact of the application.
An interesting thread on Starbucks mind hacks for customer retention:
https://twitter.com/TrungTPhan/status/1452665189208186880

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

As machine learning engineers, it is always important to justify why we choose a specific metric to evaluate the performance of our model. We need to explain why some metrics are more important than others to the problem we are analyzing. You did a good job here exploring the metrics and narrowing down on the one best suited for the problem and the dataset.
Interesting read on metrics for model evaluation:
https://www.math.ucdavis.edu/~saito/data/roc/ferri-class-perf-metrics.pdf

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.

A high level problem statement with a strategy to solve it is clearly defined.

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 of the data or input that need to be addressed have been identified.

Very nice job describing your dataset. Glad that you show some descriptive stats, show a sample of your data, go into a bit of detail in the features here and the distribution of the target variable. As this allows the reader to get an understanding of the structure of the data you are working with.
Dataset exploration: https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

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 visualizations themselves are clean and well presented, with appropriate labels and identifiers, and the right visual encoding for each data type. Visualizations are one of the best ways of transferring complex information in a condensed and easy to understand form to your reader.

An interesting read I found recently: A Visual Explanation of Gradient Descent Methods (Momentum, AdaGrad, RMSProp, Adam) (https://towardsdatascience.com/a-visual-explanation-of-gradient-descent-methods-momentum-adagrad-rmsprop-adam-f898b102325c)

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

Excellent job describing your main models here. Exploring various models and building the intuition to select the right model based on the characteristics of the dataset is one the main skills of an ML engineer. I find this reference helpful while choosing algorithms
Additionally the purpose of taking this theoretical perspective also helps your reader to understand how it's treating the data, and therefore gives more objective reasoning for why it may be optimal in this situation.

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

Very nice that you have chosen a very specific benchmark models closely related to the problem you have in your hands.!
Some interesting posts on the topic:
https://medium.com/levvel-consulting/define-benchmark-deploy-6a8d0fb0decd
https://towardsdatascience.com/benchmarking-simple-machine-learning-models-with-featureextraction-against-modern-black-box-80af734b31cc
Importance of the benchmark in machine learning work:
https://blog.dominodatalab.com/benchmarking-predictive-models/

Methodology

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.

As machine learning engineers, our primary focus should certainly be on building the best feature set we can for the algorithms we use. This is really the most "human" aspect of the ML pipeline. So it's nice to see that your improvements have a particular focus on improving the information in the feature set

- A randomized search can be a great way to search a large parameter space. It's https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
- There's a fairly famous paper that even demonstrated it would only take 60 iterations to find a fairly optimal configuration of parameters: https://www.jmlr.org/papers/v13/bergstra12a.html

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.

The goal of this section is to make our work as reproducible as possible; for any future researchers that read your work and wish to expand on it, they'll have to start by re-implementing what you have done, and they can only do that if your explanation of your work through this report is detailed and accurate. You've certainly met that requirement with your discussion here

Another idea would be to check out using Cyclical Learning Rates for Training Neural Networks(https://arxiv.org/abs/1506.01186). This is where we simply keep increasing the learning rate from a very small value, until the loss stops decreasing and then bump it up once more. We can plot the learning rate across batches to see what this looks like.

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

This is a solid step by step overview of the work required to prepare your data for proper training, and again it's written in a way that's clear and detailed https://cloud.google.com/solutions/machine-learning/data-preprocessing-for-ml-with-tf-transform-pt1

Another algorithm that is getting attention is GANs for generating new content from existing datasets https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29

Results

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.

Excellent job! Your final model is well-evaluated.
Check out this interesting book that shows how to understand the feature importance and accumulated local effects and explaining individual predictions: https://christophm.github.io/interpretable-ml-book/.

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.

Great overview of model parameters and the robustness of solution

- Interpreting model results
  SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions see the following link:
  https://github.com/slundberg/shap
- Data visualization and clustering:
  https://hypertools.readthedocs.io/en/latest/
- Cross-validation
  One can use cross validation to check the robustness of the solution. Take a look here:
  https://machinelearningmastery.com/k-fold-cross-validation/

下载项目

返回 PATH