UDACITY

< Return to Classroom

Data Scientist Capstone

REVIEW

CODE REVIEW

HISTORY

Meets Specifications

Awesome 💯 💯

Congratulations on finishing your project on time. 🏆🏆

I'm happy to say that you've exceeded all expectations and standards for the Data Scientist Capstone project. Your project showcases a remarkable level of proficiency. Your ability to effectively communicate and extract meaningful insights demonstrates your strong analytical skills

Furthermore, your code is well-structured, readable, and appropriately commented, which enhances the overall quality of your project. You have also provided clear explanations and documentation, making it easy to follow your thought process and understand your approach.

Useful References:

- How to present machine learning model performance as actionable insights to Business. [Article]
- Top 5 Machine Learning Practices Recommended by Experts [Article]

Again, congrats on your outstanding achievement! Your perseverance and dedication have paid off, and you will unquestionably be successful in all of your future pursuits. I applaud your efforts and wish you success in all of your next undertakings, both academic and professional.

For any technical assistance, you can use Knowledge Portal as well.
Have a great day 🏆

PLEASE CONSIDER RATING MY EFFORTS AS A PROJECT REVIEWER! YOUR ADVICE IS MUCH VALUED AND APPRECIATED.

Project Definition

Student provides a high-level overview of the project. Background information such as the problem domain, the project origin, and related data sets or input data is provided.

✅ High-level overview of the project

Nice! You have done an excellent job describing the overview of your project.

The high-level overview of the project is well-presented, offering a comprehensive understanding of the purpose, objectives, and methodology of the dog classifier project. The description provides a clear outline of the project's scope. The overview demonstrates a strong foundation, setting the stage for a detailed and focused analysis. Great job on delivering an informative and well-structured summary that effectively communicates the project's core aspects.

Useful References:

How to Create the Perfect Project Overview [Article]

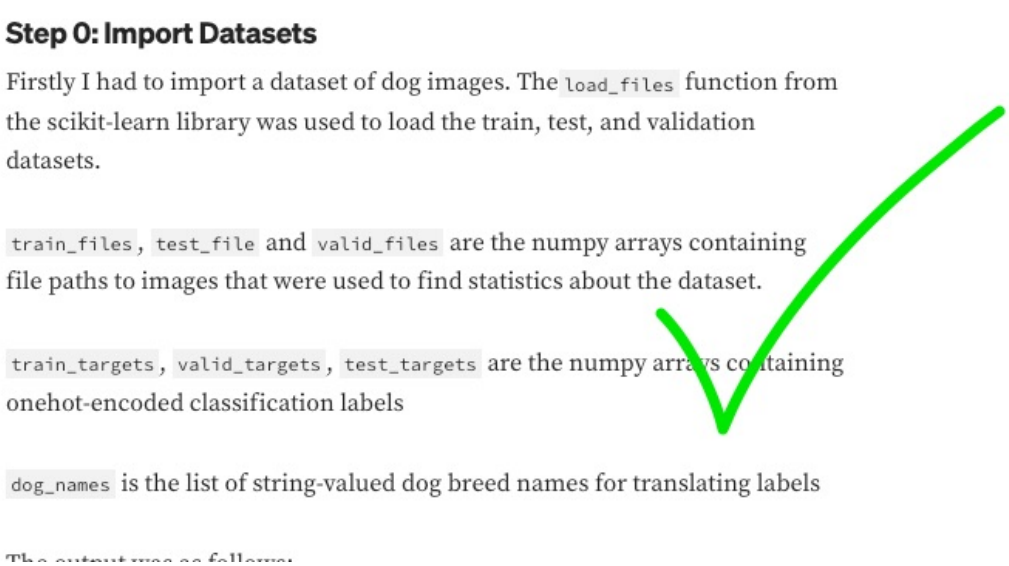
✅ Problem Domain

Good Job! 💯

The description of the problem domain is well-executed, showcasing a solid understanding of the relevant concepts and challenges. The explanation provides sufficient context and background information to grasp the significance and relevance of the dog classification problem. It demonstrates a comprehensive analysis of the problem domain, including key characteristics and potential implications.

✅ Related data sets or Input data.

Excellent! you have nicely described your dataset.



Step 0: Import Datasets

Firstly I had to import a dataset of dog images. The `load_files` function from the scikit-learn library was used to load the train, test, and validation datasets.

`train_files`, `test_file` and `valid_files` are the numpy arrays containing file paths to images that were used to find statistics about the dataset.

`train_targets`, `valid_targets`, `test_targets` are the numpy arrays containing onehot-encoded classification labels

`dog_names` is the list of string-valued dog breed names for translating labels

The output was as follows:

The description of the related data sets or input data is well-presented, providing a comprehensive overview of the sources, types, and characteristics used in the dog classifier project. The explanation includes relevant information about the data collection process, data quality, and any preprocessing steps taken. It demonstrates a clear understanding of the suitability and reliability of the data for the project.

Useful References:

17 Strategies for Dealing with Data, Big Data, and Even Bigger Data [Article]

5 Steps for How to Better Manage Your Data [Article]

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 problem which needs to be solved is clearly defined

Well done!

The problem statement is well-crafted, clearly defining the problem that needs to be solved in the dog classifier project. It effectively outlines the specific challenges and objectives, providing a clear understanding of the problem at hand. You have successfully presented a concise and well-defined problem statement, setting a solid foundation for the project

---

✅ A strategy for solving the problem, including discussing the expected solution, has been made.

Great job in formulating a robust strategy and providing clarity on the path forward!

In order to complete this project, the content of the Udacity Convolutional Neural Networks elective had to be additionally studied. The Jupyter notebook that housed the classification algorithm has the following sections which summarises the strategy in developing the algorithm for solving the above problem:-

- Step 0: Import Datasets
- Step 1: Detect Humans
- Step 2: Detect Dogs
- Step 3: Create a CNN to Classify Dog Breeds (from Scratch)
- Step 4: Use a CNN to Classify Dog Breeds (using Transfer Learning)
- Step 5: Create a CNN to Classify Dog Breeds (using Transfer Learning)
- Step 6: Write your Algorithm
- Step 7: Test Your Algorithm

The strategy for solving the problem is well-developed and clearly articulated. You have provided a comprehensive approach for tackling the problem. The strategy addresses potential challenges and considers alternative approaches, providing a well-rounded plan for implementation. You have effectively outlined the steps, techniques, and tools to be employed in solving the problem, while also discussing the expected solution in detail.

Useful References:

Different Types of CNN Architectures Explained: Examples [Article]

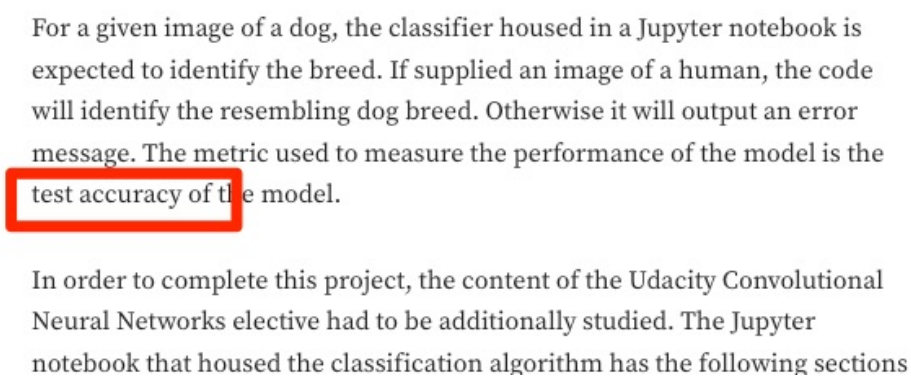
Top 5 Machine Learning Practices Recommended by Experts [Article]

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

For example, explain why you want to use the accuracy score and/or F-score to measure your model performance in a classification problem,

🔁 Metrics are justified based on the characteristics of the problem.

The justification for the chosen metrics is missing.


For a given image of a dog, the classifier housed in a Jupyter notebook is expected to identify the breed. If supplied an image of a human, the code will identify the resembling dog breed. Otherwise it will output an error message. The metric used to measure the performance of the model is the test accuracy of the model.

In order to complete this project, the content of the Udacity Convolutional Neural Networks elective had to be additionally studied. The Jupyter notebook that housed the classification algorithm has the following sections

It is crucial to provide a clear rationale for the selection of metrics, demonstrating a deep understanding of the problem and its unique challenges. Justifying metrics based on the problem characteristics helps to ensure that the chosen metrics align with the project goals and provide meaningful insights.

To improve this aspect,

it is recommended to thoroughly analyze the problem and its specific requirements. Consider factors such as the nature of the data, the objective of the analysis, and the desired outcomes. Based on these considerations, identify metrics that are relevant, appropriate, and align with the problem's characteristics.

Brief Notes:

- **Accuracy**: It refers to how close a result or prediction is to the true or correct value. It is a measure of correctness or the level of agreement between a model's output and the actual expected outcome
- **F1 Score**: The F1 score is a combination of precision and recall. It provides a balanced measure of the model's performance by considering both precision and recall at the same time. The F1 score is useful when we want to find a balance between precision and recall. It is calculated as the harmonic mean of precision and recall.
- **Precision**: Precision tells us how many of the positive predictions made by the model are actually correct. In other words, it measures the accuracy of the model when it predicts something as positive. A high precision means that the model is making fewer false positive errors.
- **Recall**: Recall tells us how many of the actual positive instances in the data were correctly predicted as positive by the model. It measures the model's ability to find all the positive instances. A high recall means that the model is making fewer false negative errors.

References:

Data Science in Medicine — Precision & Recall or Specificity & Sensitivity? [Article]

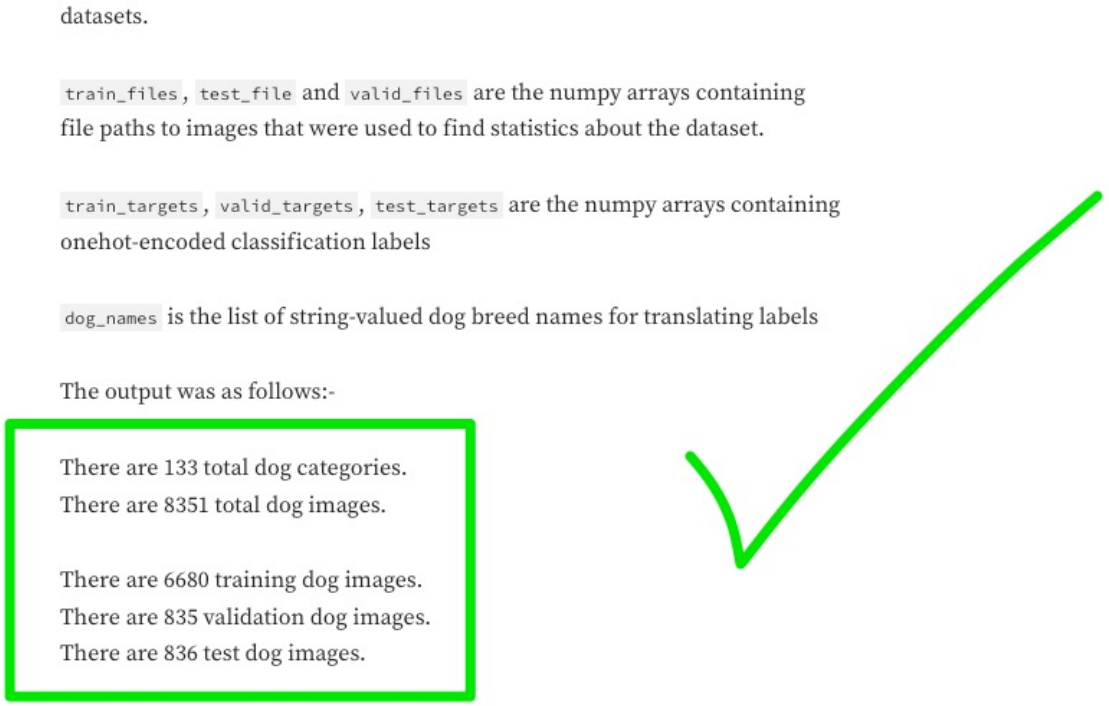
Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on? [Article]

Analysis

Features and calculated statistics relevant to the problem have been reported and discussed related to the dataset, and 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.

✅ Features and calculated statistics relevant to the problem have been reported and discussed related to the dataset, and a thorough description of the input space or input data has been made.

You have done an excellent job.


datasets.

`train_files`, `test_file` and `valid_files` are the numpy arrays containing file paths to images that were used to find statistics about the dataset.

`train_targets`, `valid_targets`, `test_targets` are the numpy arrays containing onehot-encoded classification labels

`dog_names` is the list of string-valued dog breed names for translating labels

The output was as follows:-

There are 133 total dog categories.
There are 8351 total dog images.

There are 6680 training dog images.
There are 835 validation dog images.
There are 836 test dog images.

The report provides a thorough and insightful analysis of the features and calculated statistics relevant to the problem, within the context of the image dataset used in the dog classifier project. The description of the input space or input data is comprehensive, including important image-specific attributes such as pixel dimensions, color channels, and image resolutions.

Useful References:

Handling Missing Data in Python: Causes and Solutions [Article]

✅ Abnormalities or characteristics about the data or input that need to be addressed have been identified.

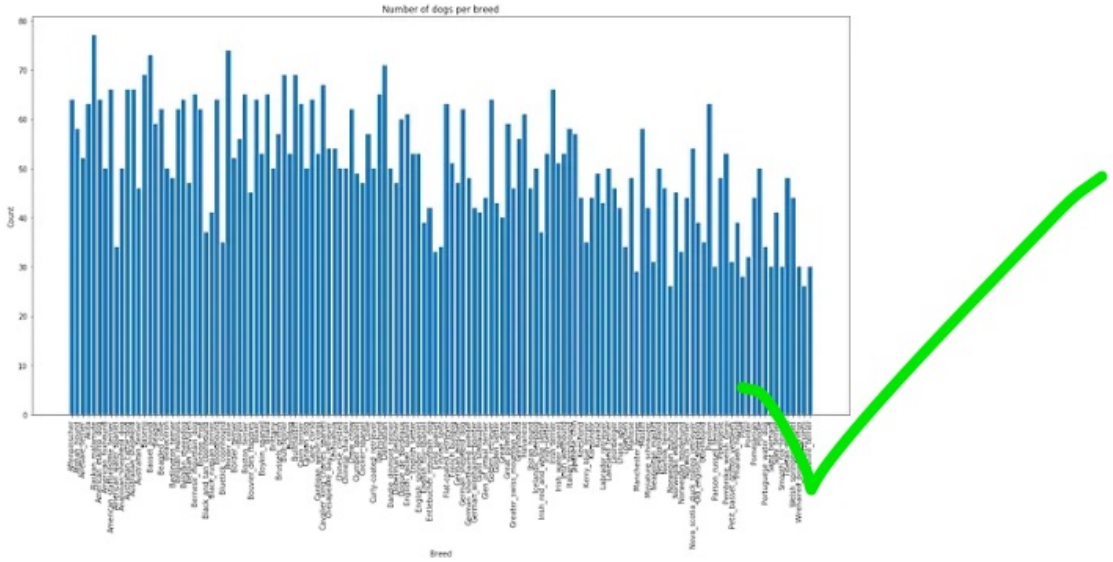
You have discussed all the abnormalities present in the data.

You have successfully explored what may have become potential challenges, such as class imbalances, noisy or corrupted images, and variations in lighting conditions. By acknowledging these factors, the report showcases a thorough understanding of the dataset's nuances.

Build data visualizations to further convey the information associated with your data exploration journey. Ensure that visualizations are appropriate for the data values you are plotting.

✅ Build data visualizations to further convey the information associated with your data exploration journey.

You have effectively used the Bar plot to extract valuable information.



The student has skillfully employed visualization techniques to showcase important aspects such as **class distributions, image samples, and feature distributions**.

These visualizations provide a clear and intuitive representation of the dataset, enabling better understanding and interpretation of the underlying patterns and characteristics

---

✅ Ensure that visualizations are appropriate for the data values you are plotting.

You have successfully ensured that the visualizations in your project are appropriate for the data values being plotted.

The choice of visualizations demonstrates a clear understanding of the data and its characteristics, resulting in accurate and meaningful representations.

You can further dive into an interactive visualization. It allows you to interact with the data in a way that traditional static visualizations do not, allowing you to dynamically explore the data and make correlations or discover patterns that might not be immediately obvious.

Useful References:

How to make great-looking, fully-interactive plots with a single line of Python [Article]

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.

✅ All preprocessing steps have been clearly documented.

You have done an excellent job in documenting all preprocessing steps in a clear and comprehensive manner.

Each step is thoroughly explained, providing a comprehensive account of the transformations, enhancements, and manipulations applied to the images prior to model training. The documentation includes specific details on techniques such as resizing, normalization, data augmentation, or noise removal, showcasing a comprehensive understanding of the preprocessing pipeline. This level of transparency and thorough documentation ensures the replicability of the results and allows for better comprehension and evaluation of the preprocessing methods employed.

✅ Abnormalities or characteristics about the data or input that needed to be addressed have been corrected.

Well Done 👍

Through careful investigation and analysis, you have identified abnormalities or characteristics in the data that could potentially impact the validity of the results. By taking appropriate corrective measures, you have mitigated the potential biases or inaccuracies caused by these issues. Your commitment to correcting these abnormalities or characteristics showcases your attention to detail and commitment to producing reliable results.

Useful References:

Image Processing in Python: Algorithms, Tools, and Methods You Should Know [Article]

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 process for which metrics, algorithms, and techniques were implemented with the given datasets or input data has been thoroughly documented

A solid step-by-step process is documented here. It is quite clear how you approached this problem. Your results would be replicable.

The documentation includes detailed descriptions of the architectures and techniques employed. You have provided insights into the rationale behind their selection, their underlying principles, and any specific parameter settings utilized. This level of detail allows others to understand and replicate the implementation process accurately.

Additionally, any preprocessing steps or data transformations performed prior to implementing the metrics, algorithms, or techniques are thoroughly documented. This includes any necessary feature engineering, scaling, encoding, or other preprocessing techniques employed.

Useful References:

ImageNet: VGGNet, ResNet, Inception, and Xception with Keras [Article]

Top 5 Machine Learning Practices Recommended by Experts [Article]

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.

✅ The process of improving upon the algorithms and techniques used is clearly documented.

You have done an excellent job in clearly documenting the improvements.

The documentation includes detailed explanations of the modifications made, optimizations applied, and the rationale behind each decision. This level of transparency allows for easy replication, evaluation, and understanding of the enhancements made to the algorithms and techniques.

Useful References:

Image Classification: Tips and Tricks From 13 Kaggle Competitions (+ Tons of References) [Article]

✅ Both the initial and final solutions are reported, along with intermediate solutions

Excellent job!

You have done an excellent job in reporting both the initial and final solutions. You have described the initial solution or baseline model employed, including its methodology, strengths, weaknesses, and initial results. This provides a solid foundation for readers to understand the starting point of your project. This allows readers to follow the iterative process and gain insights into the refinements made along the way.

Results

If a model is used, the following should hold: 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. For example, you can use cross-validation to find the best parameters.

Show and compare the results using different models, parameters, or techniques in tabular forms or charts.

Alternatively, a student may choose to answer questions with data visualizations or other means that don't involve machine learning if a different approach best helps them address their question(s) of interest.

✅ If a model is used, the following should hold: The final model's qualities — such as parameters — are evaluated in detail.

Awesome!

Your evaluation provides a thorough understanding of the characteristics and properties of the final model, ensuring its effectiveness and reliability.

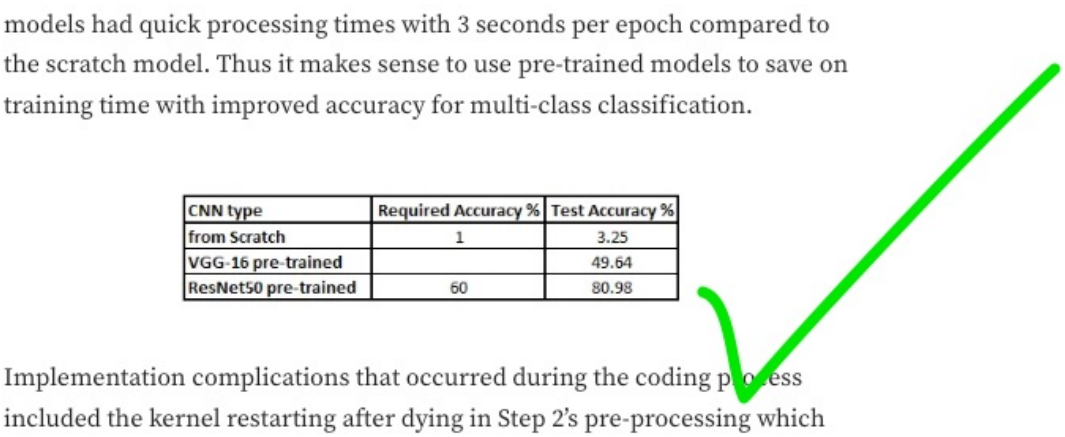
Throughout your project, you have provided a comprehensive analysis of the final model's parameters. You have discussed their values, significance, and the impact they have on the model's performance. This level of detail allows readers to gain insights into the inner workings of the model and understand the factors that contribute to its success.

References:

How to Train a Final Machine Learning Model [Article]

✅ Show and compare the results using different models, parameters, or techniques in tabular forms or charts.

Perfect! You have successfully compared the different models.


models had quick processing times with 3 seconds per epoch compared to the scratch model. Thus it makes sense to use pre-trained models to save on training time with improved accuracy for multi-class classification.

CNN type	Required Accuracy %	Test Accuracy %
from Scratch	1	3.25
VGG-16 pre-trained		49.64
ResNet50 pre-trained	60	80.98

Implementation complications that occurred during the coding process included the kernel restarting after dying in Step 2's pre-processing which

Your **table** provide a clear and concise representation of the performance and evaluation metrics associated with each approach, enabling easy comparison and analysis.

References:

Compare Machine Learning Algorithms Consistently [Article]

Graphical chart of evaluation of machine learning models [Image]

Performance comparison of all machine learning models. [Image]

The final results are discussed in detail. Explain the exploration as to why some techniques worked better than others, or how improvements were made are documented.

✅ The final results are discussed in detail

Flawless! 💯. You have nicely interpreted the results.

Your analysis and discussion provide a comprehensive understanding of the outcomes and their implications, demonstrating a thorough examination of the project's objectives and results.

References:

How to present machine learning model performance as actionable insights to Business. [Article]

✅ Explain the exploration as to why some techniques worked better than others, or how improvements were made are documented.

Wonderful! 👏

Your explanation provides valuable insights into the factors that influenced the varying effectiveness of different techniques and demonstrates a thorough understanding of the problem-solving process. By detailing the improvements, you provide transparency and allow others to understand the decision-making process involved in enhancing the techniques.

References:

Exploratory Research: What are its Method & Examples? [Article]

Conclusion

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

✅ Student adequately summarizes the end-to-end problem solution.

Superb 👍

The student has done an excellent job in summarizing the end-to-end problem solution. Their summary provides a comprehensive overview of the problem, the approach taken, and the key steps involved in solving the problem.

✅ Discusses one or two particular aspects of the project they found interesting or difficult.

Nice work discussing your final end-to-end problem solution.

Your ability to identify and discuss these points adds depth and interest to your work, enhancing the overall quality of your project. Keep up the great work!”

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

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

Good work!

By discussing potential areas for improvement and considering alternative solutions, you have displayed a proactive and analytical approach to your work. This level of critical thinking and evaluation is crucial for driving progress and ensuring continuous improvement in the implementation of the project.

Suggestion:

I would also recommend you consider potential biases and errors in the data and the model, such as overfitting, underfitting, and data leakage. These factors can significantly impact the performance and reliability of your model, and addressing them is crucial for ensuring accurate and unbiased results.

Deliverables

If the student chooses to provide a blog post the following must hold: Project report follows a well-organized structure and would be readily understood by a technical 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.

If the student chooses to submit a web-application, the following holds: There is a web application that utilizes data to inform how the web application works. The application does not need to be hosted, but directions for how to run the application on a local machine should be documented.

✅ Project report follows a well-organized structure and would be readily understood by a **technical** audience.

Your post effectively communicates your findings. 👍

This way, all the audience or stakeholders would be able to understand your insights and would be able to interpret the results visually and statistically.

Suggestion:

If removing certain technical keywords makes your sentence meaningless, then hyperlink those keywords with appropriate articles or documentation. This way, people unaware of those words can go through the documentation.

Useful References:

Art of Storytelling [Article]

✅ Each section is written in a clear, concise and specific manner

Awesome work! I really liked how you structured your report.

Writing in a clear, concise and specific manner is essential for producing effective documents. This means that each section of text should make a single point and be written in a way that is easy for the reader to understand

Suggestion:

Use shorter sentences and avoid using unnecessary words. Additionally, ensure that each sentence has a subject and verb and that the sentence's content is related to the point you are trying to make.

Student must have a Github repository of their project. The repository must have a README.md file that communicates the libraries used, the motivation for the project, the files in the repository with a small description of each, a summary of the results of the analysis, and necessary acknowledgements. If the student submits a web app rather than a blog post, then the Project Definition, Analysis, and Conclusion should be included in the README file, or in their Jupyter Notebook. Students should not use another

student's code to complete the project, but they may use other references on the web including StackOverflow and Kaggle to complete the project.

✅ Student must have a Github repository of their project

Suggestion:

Always try to keep commits as small and focussed as possible. When you are trying to fix one particular bug and you spot another one, then first resolve the first bug and commit as soon as you resolved it. Do not dive into another one without committing.

Apart from that, try to write meaningful commit messages. For example:

```
Bugfix: bugfix message
Update: update message
Correction: correction message
Added: files added or whatever
```

Best practices for using Git [Article]

✅ README file communicates the libraries used

✅ README file has motivation for the project section

✅ README file has files in the repository section

✅ README file has summary of the results of the analysis

✅ README file has acknowledgement section

README files can help potential collaborators or contributors understand the purpose and goals of the project, as well as any other relevant information

Useful References:

Manage your data science project structure in early stage [Article]

How to write a good readme for your github project? [Article]

Code is formatted neatly with comments and uses DRY principles. A README file is provided that provides. PEP8 is used as a guideline for best coding practices.

Best practices from software engineering and communication lessons are used to create a phenomenal end product that students can be proud to showcase!

✅ Code is formatted neatly with comments and uses DRY principles

✅ A README file is provided.

✅ PEP8 is used as a guideline for best coding practices.

Good Work! Your comments are constructive in understanding the flow of the analysis. 💯

Markdown text is essential in ensuring that our notebooks are easy to use and understand. The Markdown text is not a replacement for line or block comments in the code cells but rather a place to provide a broader context for the code.

You have nicely followed the pep8 style guidelines for naming your variable and functions. The use of comments throughout the code enhances its readability and understanding.

Useful References:

Python Code Quality: Tools & Best Practices [Article]

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

START