Machine Learning Specification

C00270395

Ronan Green

Contents

Introduction

This specification report includes ideas I propose for my Machine Learning portfolio. Included in this report are the sections of the portfolio, the ideas along with various possible technologies I would use and why I would use them and machine learning techniques. The appendix will have links to sites used as reference for the datasets and information on the technologies.

Portfolio Layout

In the portfolio I will include all relevant details that employers will be interested in relating my details, skills and previous projects I have worked on. The titles I plan to include in the portfolio are as follows:

Contacts

This will include my email address and phone number for employers to contact me.

About Me

This will include a small paragraph including my educational history, my main aspects of what I enjoy/good at when it comes to software development and what I can bring if I was hired(Such as personality and soft skills).

Technologies Used

This will showcase previous technologies I have used to the employer listing as languages, tools, frameworks and other relevant branches of technology which I have used.

Machine Learning Techniques Used

This will include my knowledge and experience with machine learning techniques which I have used.

Previous Projects

Building on the previous heading this will include what projects I have done and partook in and what machine learning techniques were used in this project. It will include a brief overview of what the project was about, how I approached the project, what techniques and technologies I sued, if I encountered and trouble during the project how was it overcome and my learning outcomes and benefits of the project.

Employment History

This will showcase previous relevant companies I have worked for and what college I attend for my degree showing a history of my experience.

Deployment

A few options for the Machine Learning Portfolio deployment are as follows;

- Github
- Vercel
- Netlify

Device Spec

Below is a specification of the device I will be using to run the project specification on:

- CPU: AMD Ryzen 7 8840U. This includes 8 cores, 16 Threads
- GPU: My device does not have a dedicated GPU. Its integrated GPU is an AMD Radeon 780m Graphics which has 512MB of ram available, The core count is 12 Due to this I will be making use of Google Colabs for any machine learning techniques that rely on a good GPU.
- RAM: 16GB
- Storage: 1TB

GDPR

All datasets used must comply with EU GDPR regulations, as such name will be removed and other private details that could break GDPR

Data Quality

To keep the data quality high I will be splitting the dataset into three separate percentages. 70% for the training data, 15% for the validation data and 15% for the testing data. The Training data will be used to train the model, the validation will test the model while being aware of the training data and the test data will test the model being completely isolated from all other data.

Project Ideas

Below is a list of my proposed project ideas for the Machine learning module.

Summer Olympic Athletes

Description

This project includes a comprehensive list of athletes who have participated in the summer Olympics since in inaugural year on Athens in 1896 to the 2024 Paris Olympics. The data it includes ranges from the athletes name, sex, country, year city, sport, event and medal won. This has more than 130,000 unique sets of data.

Data Application

- What this data set could be used for is predicting which country will win the most medals.
- Which country produce the most efficient athletes for certain sports.
- Countries that have been doing or worse in recent years and where they could end up in years to come.
- We can also see which country has more proficient males or more proficient females.

Annual Catch from Commonwealth Fisheries Logbooks

Description

This dataset includes annual retained catch in kilograms, by species and fishery for common wealth fisheries managed by the Australian Fisheries Management Authority. The data included in the dataset is Fishery, Gear, Year, Common_Name, Scientific_Name, CAAB(Codes for Australian Aquatic Biota), retained catch and data source.

Data Application

- This could show the total yield for certain years and expected years to come.
- What gear yields the most.
- The expected fish for what gear used and the amount.

American Real Estates sales 2001-2022

Description

This dataset is a list of all real estates sales with a sales price of $2000 or greater that occur between October 1st to September 30th of each year from 2001-2022. The dataset includes town, property address, date of sale, property type, sales prices, property assessment.

Data Application

- This can be used for estimating the sale of a certain house for a certain year.
- How much the property increases over a certain period of time.
- What months give the best ROI.
- What property types are the most valuable.

Life Expectancy Rates

Description

This dataset contains the life expectancy rates for all countries from 1960 to 2022. The dataset contains the country, the code of the country and expected rate for each year.

Data Application

- This can show the increase of each country overtime showing which countries are increasing quicker or slower.
- We could predict what countries will have the highest expectancy in years to come.
- Align it with other datasets such as wealth for countries to see how much wealth is relevant to certain life expectancy increases.

Machine Learning Techniques Used For Projects

Linear Regression

Is supervised machine learning technique. It is commonly used for sorting numerical data such as house prices or sales numbers predicting values that fall in a particular range.

Logistic Regression

Also known as logical regression is a supervised algorithm. It is used for deciding on binary data such as whether a cat is a cat or not.

Naive Bayes

This is a supervised algorithm that creates predictive models for binary or multiclassification tasks. It estimates likelihood of classification based a combined number of factors while assuming independence between the factors.

Decision Tree

Is a supervised learning algorithm that Is used for classification and predictive modelling tasks. It starts with a root node that's asks a question about he data and based on the answer the data will be directed down branches to internal nodes.

Random Forest

This combines results of multiple decision trees to get its data from the dataset.

K-Nearest-Neighbor(KNN)

Is a supervised algorithm that reflects on how close the output data is to other data points on the graph. Ths algorithm is used for classification and prediction modelling.

K-Means

Is an unsupervised algorithm used for clustering and pattern recognition. Groups data based on their proximity to one another.

Support-Vector Machine

Is a supervised algorithm used for classification and predictive modelling. It can be used with small amounts of data and works by creating a decision boundary called a hyperplane. A hyperplane is 2 separate sets of labelled data.

Technologies Used For Projects

Languages

Python

Python has become the de facto language for machine learning. This is due to many factors such as its simplicity, versatility and its wide range of libraries and frameworks.

Libraries

Python gives access to many useful libraries for machine learning. These libraries include the following:

NumPy:

NumPy is a python library that enhances arrays making them faster than regular python arrays. This library also provides functions for working with linear algebra, fourier

transform and matrices. NumPy also supports a wide range.of hardware and computing platforms. Due to it being a python library the syntax is high level accessible and productive for all level programmers.

Pandas:

Pandas is a python library used for working with datasets. It can analyse the data and has functions for cleaning exploring and manipulating data. Panda is able to get a lot of information form datasets and gives you answers about the correlation between data, average values, and making this data readable and relevant. Pandas is suited for working with spreadsheets or SQL tables. Pandas is built on top of the NumPy library meaning there is cross over in the structures used. The data used can be used in plotting functions for Matplotlib, statical analysis in SciPy and ML algorithms in Scikit-learn. It can also handle missing data, data visualization and can perform split-apply-combine operations on data sets.

Matplotlib:

Matplotlib is a python library that can be used for creating static, animated or interactive visualizations. It can generate a wide range of graphs such as line, scatter, bar, histograms, pie charts and more. Along with supplying the tools for displaying graphs it allows customization of these graphs. These graphs are of professional quality. It can work seamlessly with NumPy allowing data from arrays to be plotted easily. It also works with Seaborn, Pandas and basemap.

Seaborn:

Seaborn is a python library that builds on top of Matplotlib and closely with Pandas to make statistical graphics in python. Seaborn uses plotting functions to operate on data frames and arrays containing whole datasets. The dataset-orientated API lets the focus be on different elements of the plot and not the details on drawing the plot. Seaborn provides various plots such as relational plots, categorical plots, distribution plots, regression plots, matrix plots, multi-plot grids.

Scikit-learn:

Scikit-learn is a python library that provides diverse algorithms for classification, regression, clustering, and dimensionality reduction. The library is build using other languages like SciPy and NumPy and closely connected with Pandas and Seaborn. It provides simple and efficient tools for data mining and data analysis.

Tensorflow:

TensorFlow is a python library for numerical computation using data flow graphs. It provides functions for building and training deep learning models as it facilitates the creations of computational graphs and efficient execution on a wide range of hardware

platforms. A few features TensorFlow has are Antidifferentiation, Eager execution, Distribute, Losses, Metrics, TF.nn, Optimizers. TensorFlows's APIs use Keras to allow users to make their own machine-learning models and also helps load the data to train the model.

Keras:

Keras is an API that runs on top TensorFlow. It is a high-level API used for training and building neural networks. Its allows you to, with minimal code, to build, train and deploy deep learning models. It is known for its user friendly interface that allows all level of coders to have viable access to the API. Keras has good Extensibility and Customizability providing help for creating custom layers, loss functions and preprocessing task. Allows for complex architectures and provides subclassing to write models from scratch. With it providing sequential along with functional APIs it allows ease of use working with single input and output models as well as multiple input and output models.

Pytorch:

Pytorch is a python library that provides many convenient tools to help build neural networks and train them efficiently. PyTorch is built using tensor which is similar to NumPy. It uses the autograd library for automatic differentiation, this computes gradients of tensor. The graphs are made during run-time allowing for dynamic changes to graphs to be made.

SciPy

SciPy is a python library that is used for scientific computation and uses NumPy. It provides utility functions for optimisation, stats and signal processing. SciPy is very similar to NumPy but has added functions and is optimised. The vast amount of algorithms it provides ranges from optimisation and integration to algebraic and differential equations along with many more.

Tools

Jupyter Notebook:

Jupyter notebook is an web application used to create and share documents that contain live code, visualizations and text. It is commonly used with data science and machine learning.

Google Colab:

Google Colab is like google docs but for python code. It is a cloud based service that allows the user to write and run code in a jupyter notebook environment. Google colab

allows the user to take advantage of powerful CPUs and GPUs without having the hardware yourself. It can be used to write and execute code, develop models, and collaborate with other devs.

RapidMiner:

Rapidminer is a comprehensive data science platform with visual workflow design and full automation. It is one of the most popular data science tools. RapidMiner is used for data extraction, data mining, deep learning, machine learning, and predictive analysis.

Weka 3:

Weka Is used to provide a comprehensive suite.of tools for data analysis and predictive modelling. It helps users to analyse large datasets and applies various machine learning algorithms for task such as clustering, classification, regression, association rule data mining and data processing.

Anacoda:

Anaconda is a distribution of the python and r programming languages for scientific computing that aims to simplify package management and deployment. Packages in Anaconda are managed by the packet management system Conda.

MLFlow:

MLFlow provides an array of tools that are aimed to simplify the ML workflow. The functionalities of MLflow are rooted in several components such as Tracking, Model Registry, MLflow deployments for LLMs, Evaluate, Prompt Engineering UI, Recipes and Projects.

Database

PostgreSQL

For the database is will be using PostgreSQL. This is one of the best object-relational database management systems and is open source.

Appendix

Dataset Links

Summer Olympics

https://www.kaggle.com/datasets/stefanydeoliveira/summer-olympics-medals-1896-2024

Annual Catch from Commonwealth Fisheries Logbooks

https://data.gov.au/data/dataset/reported-retained-annual-catch-from-commonwealth-fisheries-logbooks/resource/43cb5346-e932-4e8f-8a9c-dab32dd84253

American Real Estates sales 2001-2022

https://catalog.data.gov/dataset/real-estate-sales-2001-2018

Life Expectancy Rates

https://data.worldbank.org/indicator/SP.DYN.LE00.IN?locations=IL

Machine Learning

https://www.coursera.org/articles/machine-learning-algorithms

Technology Resources

Python

https://pythonbasics.org/why-python-for-machine-learning/

NumPy

https://numpy.org/

https://www.w3schools.com/python/numpy/numpy_intro.asp

Pandas

https://www.geeksforgeeks.org/introduction-to-pandas-in-python/

https://www.w3schools.com/python/pandas/pandas_intro.asp

Matplotlib

https://matplotlib.org/

https://www.geeksforgeeks.org/python-introduction-matplotlib/

Seaborn

https://seaborn.pydata.org/tutorial/introduction

https://www.geeksforgeeks.org/introduction-to-seaborn-python/

Scikit-learn

https://datagy.io/python-scikit-learn-introduction/

https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/#features-of-scikitlearn

Tensorflow/Keras

https://www.geeksforgeeks.org/introduction-to-tensorflow/

https://www.geeksforgeeks.org/what-is-keras/

Pytorch

https://www.geeksforgeeks.org/getting-started-with-pytorch/

https://builtin.com/machine-learning/pytorch

SciPy

https://scipy.org/

https://www.w3schools.com/python/scipy/scipy_intro.php

Jupyter Notebook

https://realpython.com/jupyter-notebook-introduction/

Google Colab

https://bytexd.com/what-is-google-colab-a-beginner-guide/

RapidMiner

https://mindmajix.com/rapidminer-tutorial

https://www.analyticsvidhya.com/blog/2021/10/intro-to-rapidminer-a-no-code-development-platform-for-data-mining-with-case-study/

Weka 3

https://www.geeksforgeeks.org/introduction-to-weka-key-features-and-applications/

Anacoda

https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)

MLFlow

https://mlflow.org/docs/latest/introduction/index.html

PostgreSQL

https://www.geeksforgeeks.org/what-is-postgresql-introduction/