Youssef Hosni

DATA SCIENCE

PORTFOLIO FOR SUCCESS

Data SCIENCE PORTFOLIO FOR SUCCESS

Ultimate Guide to Building a Data Science Portfolio

YOUSSEF HOSNI

For my father, mother and my wife who supported me in every move. My father, I hope you are in a better place.

~ Youssef

Table of Contents

Introduction

In the era of data-driven decision-making, where information reigns supreme, the field of **Data Science** stands as an indispensable force shaping industries, economies, and even our daily lives. As organizations harness the power of data to gain insights, solve complex problems, and fuel innovation, the demand for skilled data scientists has reached unprecedented heights. Whether you're a seasoned data professional or just embarking on your journey into the world of data science, one critical aspect remains constant: the need for a robust and compelling portfolio.

Welcome to "Data Science Portfolio for Success: The Ultimate Guide on Building Data Science Portfolio." In this transformative book, we embark on a journey that will empower you to not only navigate the intricate terrain of data science but also master the art of showcasing your skills and expertise. Whether you aim to land your dream job, advance your career, or contribute to groundbreaking research, your data science portfolio will be your compass, guiding you toward your objectives.

In the pages that follow, we will delve deep into the world of data science portfolio starting with the importance of data science portfolio, followed by a step-by-step guide to build your portfolio and industry level projects. After that you will be introduced to tools to host your portfolio and where to find datasets for your portfolio projects. Finally, we conclude with the mistakes that you should avoid when building your data science portfolio.

Prepare to embark on a transformative journey that will not only elevate your data science skills but also set you on a trajectory toward a successful and impactful career. Let's get started on your quest to master the art of building a data science portfolio that unlocks doors to endless possibilities.

1. The Importance of Having a Portfolio as a Data Scientist

In today’s competitive job market, simply having a degree in data science may not be enough to land your dream job as a data scientist. Employers are increasingly looking for candidates who can demonstrate their skills and expertise in the field through real-world projects and experience.

This is where having a strong data science portfolio comes into play. In this chapter, we will discuss what a data science portfolio is, why it is important, and how you can build a solid portfolio that will showcase your abilities and make you stand out in the job market.

1.1. What is a Data Science Portfolio?

A Data Science Portfolio is a collection of projects, code, and other artifacts that showcase a data scientist’s skills and capabilities to potential employers or collaborators. It typically includes real-world data analyses, machine learning models, data visualizations, and other examples of work that demonstrate a data scientist’s ability to manipulate and extract insights from data.

A well-designed portfolio should provide a clear and compelling narrative of the data scientist’s skills, interests, and professional goals, and demonstrate their ability to apply data science techniques to real-world problems. This can be achieved by including a diverse range of projects that showcase different aspects of the data scientist’s skill set, such as data cleaning, feature engineering, model selection, and evaluation.

In addition to demonstrating technical skills, a strong data science portfolio should also highlight a data scientist’s ability to communicate their findings to non-technical stakeholders, through clear and concise visualizations, reports, and presentations. Overall, a data science portfolio is an essential tool for any data scientist looking to showcase their work and advance their career in the field.

1.2. Why Is Having a Solid Portfolio Important?

Having a solid portfolio is essential for anyone looking to establish themselves as a skilled and credible professional in their field. This is especially true for data scientists, as the field is highly competitive and rapidly evolving, and employers are always looking for evidence of an applicant’s abilities and experience.

Here are a few key reasons why having a solid portfolio is important for data scientists:

- **Demonstrates Your Skills and Abilities**: A portfolio provides a tangible demonstration of your skills and abilities as a data scientist. It shows potential employers or clients what you are capable of and provides concrete evidence of your experience and expertise. This can help you stand out from other applicants and increase your chances of being hired or contracted.
- **Showcases Your Creativity and Problem-Solving Abilities**: A portfolio allows you to showcase your creativity and problem-solving abilities. By sharing examples of your work, you can demonstrate your ability to think critically and creatively about complex data problems and come up with innovative solutions.
- **Provides Proof of Concept**: A portfolio provides proof of concept that you can apply your skills to real-world problems. Employers want to see that you can use your skills to solve actual business problems, and a portfolio provides a powerful way to demonstrate this.
- **Demonstrates Your Communication and Collaboration Skills**: A strong portfolio should also showcase your communication and collaboration skills. As a data scientist, you will need to work closely with stakeholders who may not have technical backgrounds, and your ability to communicate your findings effectively will be critical. A portfolio that includes well-designed visualizations and clear explanations of your

work demonstrates your ability to communicate technical information to non-technical audiences.

- **Discussion Topic**: During an interview, a portfolio might be used as a discussion starter. Applicants may go over their projects, difficulties encountered, and solutions they came up with. By doing so, the applicant can establish a relationship with the interviewer and show that they can clearly and succinctly express complicated ideas.

1.3. How to Build a Strong Portfolio?

Building a strong data science portfolio is important to showcase your skills, knowledge, and experience to potential employers or clients.

Here are some tips to help you create a strong data science portfolio:

- **Start with a clear focus:** Identity what kind of data science projects you want to work on and what areas of data science you are most interested in. This will help you tailor your portfolio to your goals and target audience.
- **Include a variety of projects:** Include a mix of projects that showcase different skills, such as data cleaning and preprocessing, exploratory data analysis, machine learning, and data visualization.

- **Choose relevant unique datasets:** Use datasets that are relevant to your target audience and your area of interest. It’s also a good idea to choose datasets that are publicly available so that others can easily reproduce your results. However, try to stay away from the very popular overused datasets.
- **Provide context:** For each project, provide context about the problem you were trying to solve, the methods you used, and the results you achieved. This will help others understand your thought process and the impact of your work.
- **Make it visually appealing:** Use clear and concise visuals to showcase your results. Use graphs, charts, and other visualizations to help tell the story of your work.
- **Use industry-standard tools and techniques:** Use industry-standard tools and techniques to showcase your skills and knowledge. This will help potential employers or clients understand your familiarity with the tools they use.
- **Keep it up to date:** As you complete new projects, continue to update your portfolio to showcase your most recent work.

1.4. What You Should Include in Your Portfolio?

1.4.1 Three Types of Data Science Projects

- **Data Analysis Projects:** The first project you should have on your portfolio is a data analysis project. In this project, you will focus on important data-related skills such as data cleaning, data collection, data exploration, data visualization and storytelling. Data analysis projects will demonstrate your ability to extract insights from data and persuade others. This has a large impact on the business value you can deliver and is an important piece of your portfolio.
- **Machine Learning Projects:** The second type of project should be a machine learning project that covers basic machine learning tasks. This includes having at least three machine learning projects cover regression, classification, and clustering tasks. It will be very helpful if you can work with different data types, such as tabular data, time series data, images, videos, and text. For each data type, this will require different techniques, skills, and tools, which will show up your skills and make you stand out.
- **End-to-End Projects:** Having an end-to-end data science project will show that you're capable of building systems that are customer-facing. Customer-facing systems involve high-

performance code that can be run multiple times with different pieces of data to generate different outputs. An example is a system that predicts the stock market — it will download new market data every morning, then predict which stocks will do well during the day. In order to show we can build operational systems; we'll need to build an end-to-end project. An end-to-end project takes in and processes data and then generates some output. Often, this is the result of a machine learning algorithm, but it can also be another output, like the total number of rows matching certain criteria.

1.4.2. Comprehensive Documentation

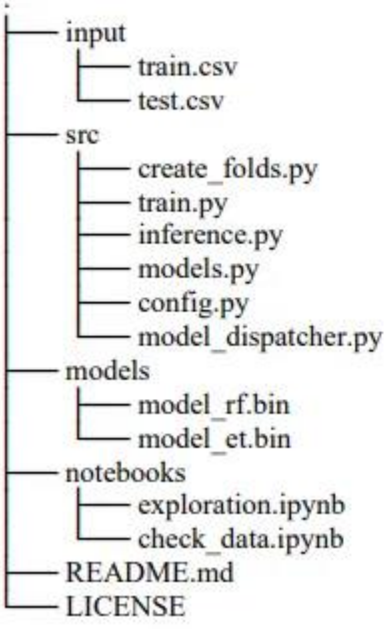
It is important to have comprehensive documentation for each project to be able to show your work in a proper way, else you will not be able to show your work and results. A comprehensive project readme should have the following sections, according to **Andrew Jones**:

- **Project Overview**: First, you should provide the project background and motivation and what is the goal of the project.
- **Data Overview & Preparation Steps**: A brief overview of the data used, how you collected the data, and how you prepared it, which includes data preprocessing and feature engineering. Also, it is also important to provide a summary of the data.

- **Methodology Overview**: You should also mention the techniques and tools used to achieve the project goals.
- **Results & Outcomes**: Finally, you should also mention the results and the project outcomes and what they mean.
- **Growth & Next Steps**: You should also mention what are the next steps and what you would do if you had more time.

1.4.3. Solid Project Structure

It is important to have a solid structure for your project as this will show that you can work on a production-level code. A very good starting point is the project structure shown in the figure below, which was provided by Abhishek Thakur in his book Approaching (Almost) Any Machine Learning Problem.



```
.
├── input
│   ├── train.csv
│   └── test.csv
├── src
│   ├── create_folds.py
│   ├── train.py
│   ├── inference.py
│   ├── models.py
│   ├── config.py
│   └── model_dispatcher.py
├── models
│   ├── model_rf.bin
│   └── model_et.bin
├── notebooks
│   ├── exploration.ipynb
│   └── check_data.ipynb
├── README.md
└── LICENSE
```

Figure 1. Machine learning project structure

- **input/:** This folder consists of all the input files and data for your machine learning project. If you are working on NLP projects, you can keep your embeddings here. If you are working on image projects, all images go to a subfolder inside this folder.
- **src/:** We will keep all the Python scripts associated with the project here. If I talk about a Python script, i.e., any *.py file, it is stored in the src folder.

- **models/:** This folder keeps all the trained models.
- **notebooks/:** All jupyter notebooks (i.e., any *.ipynb file) are stored in the notebooks folder.
- **README.md:** This is a markdown file where you can describe your project and write instructions on how to train the model or to serve this in a production environment.
- **LICENSE:** This is a simple text file that consists of a license for a project, such as MIT, Apache, etc.

2. How to Build a Data Science Portfolio That Will Land You a Job?

In this chapter, we will discuss the key components of a successful data science portfolio project and provide tips and best practices for building one that will make you stand out to potential employers. Whether you’re just getting started in the field or looking to take your career to the next level, this guide will help you create a portfolio that showcases your skills and experience in the best possible light.

2.1. Select a Domain of Interest

The first step to building a strong data science portfolio is to focus on a certain domain of interest. Data science and AI, at the end of the day, are tools that are used to solve a problem, improve performance, or automate a certain task. Therefore, it is important to decide which domain you would like to apply your data science skills.

This might depend on your previous experience. If you have work experience in a certain domain, it will be easier to find business problems to work on. It can also be a domain you are interested in and would like to use your skills to make an impact in this domain.

It is important to mention that the more you have experience in this domain, the more you will be able to get unique ideas, and the better your projects will be. In addition to that, it will give you a great advantage in the market and make you stand out. As you will have a very good understanding of the data collection process and what it means, it will improve your skills in engineering the features of the data.

Actions:

- Choose three domains of interest depending on your experience and research background.

- You should take into consideration your career goal and whether you would like to work in research or in industry.
- You can find more about different domains and how data science and AI are used to solve business problems in this **article**.

2.3. Prioritize Your Interest Based on the Market Demand

The next step after selecting two or three domains of interest is to prioritize and arrange them based on the market needs and demands. For example, I am living in Finland, so if I would like to join this market, I will have to do some market research to understand the market demand and know the companies working in these domains.

Based on my market research I found that there are a lot of opportunities and demands in the telecommunication domain (Nokia, Elisa, DNA), gaming domain(Unity, Rovio), and the Fintech industry. So, I should prioritize my interests based on this when I start to build an end-to-end project.

Actions:

- Select the market you would like to work in.
- Do market research and find the intersection between your interest and the market need.

- Select one domain that meets the previous criteria.
- Select the companies that are working in this intersection and are regularly hiring.

2.4. Define Important Case Studies in the Market

Now you have selected the domain of interest that meets your interest and the market demand, and you got good ideas about the companies that are hiring in this domain and the job requirements. It is time to define the case study or business problem to start building a solution to solve it or answer the business requirements.

To come up with real-world business problems or questions in this domain, you can do further research about the selected companies above and know what they are working on and what they use data science for. This can sometimes be found in the job requirements or on the company website. However, if you cannot find it there you can take a further step and start to ask one of the data scientists that are working there.

I believe this is a very useful and goal-oriented approach because not only will you be working on similar problems as your dream company is working on, but it will make it easier to get a job offer there. You will also get a good idea about the tools and the technology stack at these companies so you can learn them and use them in creating your projects.

Another approach is to interview experts from this domain and ask them about the most important questions or problems they have, and they wish they could use the data to solve it.

Actions:

- Read the recent data science job requirements for the companies you are interested in or read about the projects these companies are currently working on.
- If you do not find much information on the job requirements or the company website, you can contact data scientists that are working there.
- Find three case studies or business problems they are using data science to solve it.
- Repeat this for different companies you are interested in working there.

2.5. Choose Different Case Studies

Now you have multiple case studies to work on in your domain of interest. It is time now to narrow down your selection. My suggestion is to have at least three solid case studies that cover the basic machine learning tasks, which is regression, classification, and clustering.

In addition to that, try to focus on having them solved using different data types, and you narrow your selection and focus more on the data types you are passionate about. For example, if you would like to show your computer vision skills, you can focus more on case studies that need computer vision skills.

To elaborate more, let's take a practical example. Let's assume that you are interested in working in the healthcare domain. So, after searching, you came up with multiple case studies in this field, and you are interested in the computer vision domain. Here are three case studies that cover different machine learning tasks and are solved using computer vision skills:

- **Regression:** Movement disorder detection using pose estimation.
- **Classification**: Brain tumor detection and classification
- **Clustering**: Alzheimer's disease analysts using clustering

Actions:

- Discover the different areas of AI and know your interests.
- Narrow down your case studies into at least three that cover the basic machine learning tasks and focus on your area of interest in AI.
- Define which case study is solved by which machine learning task and which data type.

2.6. Brainstorm Data Science Solutions

Now you have defined the case study you would like to work on and defined the business questions for this problem. Do not rush into building the first solution that comes into your mind. Instead, take your time to brainstorm different potential solutions for it and study each of them and see which one will lead to better results and meet your learning goals and set of skills.

After brainstorming the solutions and choosing the suitable one you will need to assess the feasibility and value of potential solutions. This can be done by reviewing the published research papers on this topic or you can discuss with an expert your potential solution and see whether it is reasonable and will achieve the expected results or not. This is also a very critical step as it will save you a lot of time, effort, and future disappointments.

Actions:

- Brainstorm different AI& data science solutions for your business problem
- Evaluate them based on your criteria and select the one that best meets them.
- Validate your potential solution.

2.7. Determine Success Metrics

Once you have brainstormed the potential solutions and validated their feasibility and value, it is time to determine the success metrics you aim for this project and solution.

This includes both machine learning metrics, or what are known as offline metrics, such as (accuracy and F1 score) and business metrics, or what is known as online metrics (revenue, click-through rate).

Machine learning teams are often most comfortable with metrics that a learning algorithm can optimize. But we may need to stretch outside our comfort zone to come up with business metrics, such as those related to user engagement, revenue, and so on. Unfortunately, not every business problem can be reduced to optimizing test set accuracy! If you can’t determine reasonable milestones, it may be a sign that you need to learn more about the problem.

Actions:

- Study the problem and define the success metrics for your project, both the machine learning and the business metrics.

2.8. Collect the Data

Now, as your idea is ready, it is up to me to get your hands dirty with data. You need to collect real-world data to answer your business questions or to train the models. It is very important to use a unique dataset that is representative of your problem.

Kindly stay away from well-known datasets such as the Titanic, California house prices, Iris flowers dataset, and similar well-known datasets. They are very good for beginners and for educational projects, but they will harm you if your portfolio projects are with them.

Here are some suggested ways to collect unique datasets to develop your solution based on them:

- Kaggle
- Scrape your own data.
- Ask for data.
- Use open datasets from universities, NGO organizations, or governmental organizations.

Actions:

- Search for different data sources that can provide you with a unique dataset that can fit your project.
- Collect and store the data.

2.9. Clean & Prepare the Data

Now you have the data. The next step is to make it ready for modeling. This includes cleaning the data and applying different feature engineering techniques to it to get the best out of it.

This step is very demanding, especially if your data is real-world data which might have a lot of missing data, outliers, and other defects which are very common in real-world data.

In addition to that, you will have to explore the data to have a better understanding of it and to guide you when you start to engineer its feature and make it ready for the modeling step. Feature engineering will include data preprocessing, feature selection and dimensionality reduction, and more, depending on your problem and the collected data.

Actions:

- Clean the data.
- Explore the data.
- Feature engineering to make it ready for modeling.

2.10. Train & Evaluate the Model

Now it is time to train the machine learning model using your data. This will include several steps. First, you must choose the models to use. This will depend on many factors, such as:

- Model explainability
- In memory vs. out memory
- Number of features and instances
- Categorical vs. numerical features

- Data normality
- Training speed
- Prediction speed

You can find more information in this **article**. Next, you will train and evaluate the model. This step includes splitting the data, training the model, choosing suitable and representative evaluation metrics, and hyperparameter optimization.

Actions:

- Select suitable models for your problem.
- Split the data.
- Train the model
- Choose suitable evaluation metrics.
- Optimize the model hyperparameters.
- Test your model on the testing data.

2.11. Make them End to End

The final step is to make your project an end-to-end project. Many people usually stop at the previous. By this, you miss a big chance of making your project a real project which will help you in your job searching journey and make you stand out from other candidates.

To make your project an end-to-end project, you will need to take your model a further step and deeply it into production and integrate it into a web or mobile application. After that, you will start to monitor the model and see how it responds to new data. Based on the model performance in production you will have to retrain the model, change it, collect more data, engineer the features in a different and so on.

Actions:

- Deploy the trained model into production.
- Integrate the model into a mobile or a web application.
- Monitor the model's performance.
- Iterate

2.12. Publish & Talk About It

The final step is to publish your project on your GitHub page and create a comprehensive readme file for it. Your readme file should contain this:

- Motivation & business problem statement
- How the data was collected
- Main data cleaning steps
- Comprehensive data exploration plots
- Main feature engineering steps
- The model used and why you chose it.
- Evaluation metrics and why you chose it.
- The model performance
- How to try the model in production

Having a comprehensive readme file for your project is a very step that a lot of people actually do not focus on. It will make your work and project more valuable and accessible since many people will only read the readme before deciding to go through the code. It also shows your documentation and insight communication skills which are critical skills for aspiring data scientists.

Finally, you need to start talking about your project to grab the attention and to show the people what you are capable of. You can record a short video of your project while working in real time and publish it on your social media channels, especially LinkedIn & Twitter and invite your connections to try it and give feedback on this experience.

You can also write a blog explaining each step and show the insights you got from the data. You can create a YouTube video explaining the project step and show the results and the insights you got from the data and how you answered the business questions, and how the model works in production.

Actions:

- Upload your project on GitHub & Publish it on your professional social media channel.
- Write a comprehensive readme file for your project.
- Record a short video of your project to demonstrate how it works.
- Invite people to try your project and give you feedback.
- Write a blog about your project.
- Record a long video explaining the project steps.

Working on data science and AI projects is an iterative process. If, at any step, you find that the current direction is infeasible, return to an earlier step and proceed with your new understanding. Also, if you deployed the model into production and got unexpected results, you can go back and iterate till you can get the results that meet your success criteria.

Finally, I would like to conclude with these words from Andrew NG

Avoid analysis paralysis. It doesn’t make sense to spend a month deciding whether to work on a project that would take a week to complete. You’ll work on multiple projects over the course of your career, so you’ll have ample opportunity to refine your thinking on what’s worthwhile.

3. Building Industry-Level Data Science Projects: A Step-by-Step Guide

While the demand for data scientists & AI engineers continues to rise, there remains a significant gap between academic knowledge and practical implementation, especially when it comes to building industry-level data science projects.

This chapter aims to bridge that gap by providing a comprehensive step-by-step guide to building industry-level data science projects. Whether you are a budding data scientist looking to apply your skills in a real-world setting or an industry professional seeking to leverage data science for your organization's growth, this chapter will equip you with the essential knowledge and practical strategies required to navigate the complexities of data science project development.

Throughout this chapter, we will explore the key stages involved in building industry-level data science projects, starting from project scoping and data acquisition to modeling, evaluation, and deployment. Each step will be accompanied by best practices, real-world examples, and practical tips to help you navigate the challenges and make informed decisions at every stage of the project lifecycle.

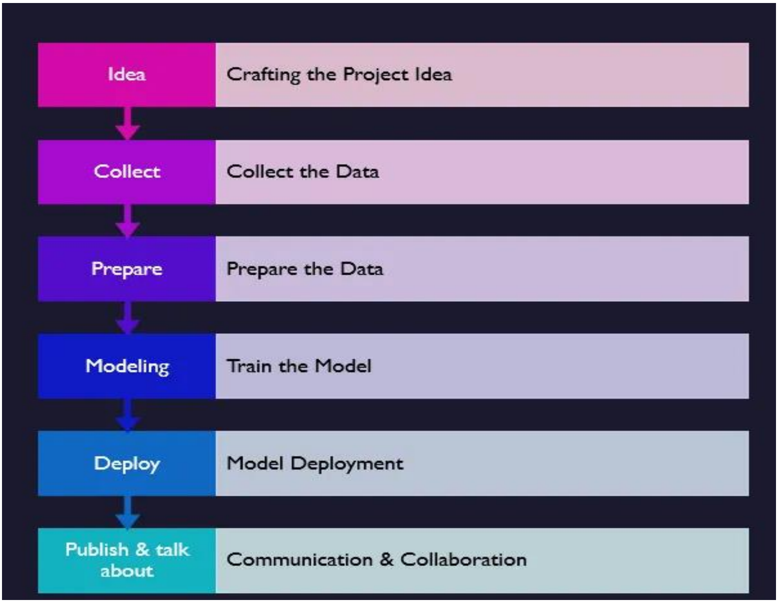


Figure 2. Building industry level data science projects: A step-by-step guide

3.1. Crafting the Project Idea

The first crucial step in building an industry-level data science project is to craft a well-defined and impactful project idea. A strong project idea serves as the foundation for your entire endeavor, driving the focus, scope, and direction of your work. In this section, we will explore the key considerations and strategies involved in crafting a compelling project idea.

3.1.1. Define Industry-Level Use Case

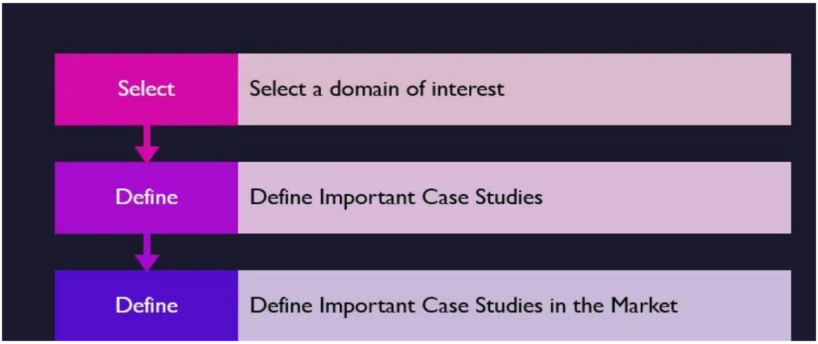

Select
Select a domain of interest
Define
Define Important Case Studies
Define
Define Important Case Studies in the Market

Figure 3. Define industry-level use case.

A. Select a domain of interest.

The first step is to define the domain of interest for your idea. Data science and AI, at the end of the day, are tools that are used to solve a problem, improve performance, or automate a certain task. Therefore, it is important to decide which domain you would like to apply your data science skills.

This might depend on your previous experience. If you have work experience in a certain domain, it will be easier to find business problems to work on. It can also be a domain you are interested in and would like to use your skills to make an impact in this domain.

It is important to mention that the more you have experience in this domain, the more you will be able to get unique ideas, and the better your projects will be. In addition to that, it will give you a great advantage in the market and make you stand out. As you will have a very good understanding of the data collection process and what it means, it will improve your skills in engineering the features of the data.

B. Prioritize your interest based on the market demand.

The next step after selecting two or three domains of interest is to prioritize your choices based on the market needs and demands. For example, I am living in Finland, so if I would like to join this market, I will have to do some market research to understand the market demand and know the companies working in these domains.

Based on my market research, I found that there are a lot of opportunities and demands in the telecommunication domain (Nokia, Elisa, DNA), gaming domain(Unity, Rovio), and the Fintech industry. So, I should prioritize my interests based on this when I start to build an end-to-end project.

C. Define important case studies in the market.

Now you have selected the domain of interest that meets your interest and the market demand, and you got good ideas about the companies that are hiring in this domain and the job requirements. It is time to define the case study or business problem to start building a solution to solve it or answer the business requirements.

To come up with real-world business problems or questions in this domain, you can do further research about the selected companies above and know what they are working on and what they use data science for. This can sometimes be found in the job requirements or on the company website. However, if you cannot find it there you can take a further step and start asking data scientists that are working there.

I believe this is a very useful and goal-oriented approach because not only will you be working on similar problems as your dream company is working on, but it will make it easier to get a job offer there. You will also get a good idea about the tools and the technology stack at these companies so you can learn them and use them in creating your projects.

Another approach is to interview experts from this domain and ask them about the most important questions or problems they have, and they wish they could use the data to solve them.

3.1.2. Setting Baseline & Define KPI

Setting a baseline and defining key performance indicators (KPIs) are crucial steps in building industry-level data science projects. These steps provide a clear starting point and a measurable framework to evaluate the success and progress of the project.

By setting a baseline, data scientists can establish a reference point against which they can compare the performance of their models and algorithms. This baseline represents the initial state or existing solution that the project aims to improve upon. You can find more information on defining a baseline for your data science & machine learning projects in this **article**.

Defining KPIs, on the other hand, enables the project team to identify and track specific metrics that align with the project's objectives and desired outcomes. KPIs can be diverse and may include measures such as accuracy, precision, recall, customer satisfaction, revenue growth, or operational efficiency. It is essential to select KPIs that are relevant, quantifiable, and directly linked to the project's goals. These indicators serve as milestones and guide the project's direction, ensuring that efforts are focused on achieving tangible and measurable results.

By establishing a baseline and defining KPIs, data science projects can operate with clarity and purpose, facilitating effective decision-making and driving meaningful progress toward their intended outcomes.

3.1.3. High-Level System Design

Now you have defined the case study you would like to work on and defined the business questions for this problem. Do not rush into building the first solution that comes into your mind. Instead, take your time to brainstorm different potential solutions for it and study each of them and see which one will lead to better results and meet your learning goals and set of skills.

After brainstorming the solutions and choosing the suitable one you will need to assess the feasibility and value of potential solutions. This can be done by reviewing the published research papers on this topic or you can discuss with an expert your potential solution and see whether it is reasonable and will achieve the expected results or not. This is also a very critical step as it will save you a lot of time, effort, and future disappointments.

3.2. Collect the Data

Now, as your idea is ready, it is up to me to get your hands dirty with data. You need to collect real-world data to answer your business questions and help you solve your problem. It is very important to use a unique dataset that is representative of your problem.

Kindly stay away from well-known datasets such as the Titanic, California house prices, Iris flowers dataset, and similar well-known datasets. They are very good for beginners and for educational projects, but they will harm you if your portfolio projects are with them.
Here are some ten resources to collect unique datasets to develop your solution based on them:

- Google Dataset Search
- Kaggle

- UCI Machine Learning Repository
- Data.gov
- Awesome Public Datasets GitHub Repo
- Reddit's /r/datasets
- Pudding.cool
- Fivethirtyeight
- KDnuggets
- Buzzfeed

You can find more information about each of them in this **article.**

3.3. Prepare the Data

Now you have the data. The next step is to make it ready for modeling. This includes cleaning the data and applying different feature engineering techniques to it to get the best out of it.

Data preprocessing is a critical step in the data analysis pipeline that involves transforming raw data into a clean, structured format suitable for further analysis and modeling. It aims to handle common data issues, such as missing values, outliers, inconsistencies, and data incompatibilities, which can adversely affect the accuracy and reliability of the results.

The following are some common techniques and steps involved in data preprocessing:

- **Data Cleaning**: This step involves handling missing values by either imputing them with appropriate values or removing the instances with missing data if they are deemed irrelevant. Outliers, which are extreme values that deviate significantly from most of the data, can be identified and treated through methods like filtering or replacing them with statistically reasonable values.
- **Data Integration**: When dealing with multiple data sources, data integration is necessary to combine and merge the data into a unified dataset. This may involve resolving schema conflicts, standardizing units of measurement, and reconciling differences in data structures.
- **Data Transformation**: Data transformation involves converting the data into a suitable format for analysis. This can include scaling numerical data to a consistent range, encoding categorical variables into numerical representations, and applying mathematical functions or transformations to achieve better distribution properties.
- **Feature Selection**: Feature selection aims to identify the most relevant and informative features or attributes that contribute significantly to the analysis or modeling task. This helps to reduce dimensionality, enhance model performance, and mitigate the curse of dimensionality.

- **Discretization**: Discretization is the process of transforming continuous variables into categorical or ordinal variables. This can be useful in cases where the relationships between variables are non-linear or when certain algorithms require categorical inputs.
- **Data Normalization/Standardization**: Normalizing or standardizing the data ensures that all variables are on a similar scale, preventing certain features from dominating the analysis due to their larger magnitude. Normalization typically involves scaling the data to a range of 0 to 1, while standardization involves transforming the data to have a mean of 0 and a standard deviation of 1.
- **Data Sampling**: In cases where the dataset is imbalanced, with one class significantly outnumbering the others, sampling techniques such as oversampling or under sampling can be applied to balance the representation of different classes. This helps to prevent bias in subsequent modeling steps.
- **Data Splitting**: Before modeling, it is essential to split the preprocessed dataset into training, validation, and testing sets. The training set is used to build the model, the validation set is used to fine-tune model parameters and make decisions regarding model selection, and the testing set is used to evaluate the final model's performance on unseen data.

Data preprocessing is an iterative process that often requires experimentation, exploration, and domain knowledge. By carefully preparing and cleaning the data, analysts, and data scientists can enhance the quality of their analysis, improve the accuracy of models, and derive more reliable insights from the data.

3.4. Train the Models

Now it is time to train the machine learning model using your data. This will include several steps. It starts with model selection, training the model, and then hyperparameter optimization. After that, you will evaluate the model and analysis any bottleneck, and iterate. Once the model is ready you can push it and implement the inference pipeline. Let's go through each of these stages in more detail:

3.4.1. Model Selection

Model selection is a crucial step in the data science process that involves choosing the most appropriate machine learning or statistical model for a given problem. The goal of model selection is to identify the model that can best capture the patterns and relationships within the data and make accurate predictions or provide meaningful insights.

Here are some key considerations and steps involved in model selection:

- **Problem to Solve**: Clearly understand the problem you are trying to solve and the goals you want to achieve with the model. This includes determining whether it is a regression, classification, clustering, or another type of problem.
- **Data Understanding:** Gain a thorough understanding of the data, including its characteristics, size, and distribution. Assess the relationships between variables, identify potential outliers or anomalies, and consider any specific data requirements or constraints.
- **Identify Relevant Models**: Based on the problem type and data characteristics, identify a range of models that are suitable for the task. This may involve consulting literature, domain experts, or previous research in the field.
- **Evaluate Model Complexity**: Consider the trade-off between model complexity and interpretability. Simpler models may be more interpretable and easier to implement, while more complex models may offer higher predictive performance but could be harder to interpret.
- **Consider Algorithm Assumptions**: Understand the assumptions and limitations of each model algorithm. Some algorithms may assume linearity, independence, or specific distributions within the data. Make sure these assumptions align with the characteristics of your data.

You can find more information about model selection in this **article**.

3.4.2. Model Training & Hyperparameter Tuning

Model training and hyperparameter tuning are essential steps in building machine-learning models for industry-level projects. Model training involves fitting the model to the training data to learn the underlying patterns and relationships, while hyperparameter tuning involves finding the optimal values for the hyperparameters that govern the model's behavior.

Here is an overview of the process:

- **Initialize Hyperparameters**: Set initial values for the model's hyperparameters. Hyperparameters are parameters that are not learned from the data but are set prior to training and affect the model's behavior (e.g., learning rate, regularization strength, number of hidden units). For a lot of problems, you will find recommendations for hyperparameter initialization that will help your model to converge faster.
- **Train the Model**: Use the training data to fit the model and learn the optimal parameters. This involves feeding the input data through the model, calculating the model's output, comparing it with the actual output, and updating the

model's parameters using optimization algorithms like gradient descent.

- **Evaluate Model Performance**: Assess the model's performance on the validation set using suitable evaluation metrics (e.g., accuracy, precision, recall, mean squared error). This provides an indication of how well the model generalizes to unseen data.
- **Hyperparameter Tuning**: Adjust the hyperparameter values to improve the model's performance. This can be done through various methods, including grid search, random search, or more advanced techniques like Bayesian optimization or genetic algorithms. Iterate through different hyperparameter configurations and evaluate their impact on the model's performance.
- **Cross-Validation**: Perform cross-validation during the hyperparameter tuning process to obtain more reliable estimates of the model's performance. This involves repeatedly splitting the data into training and validation sets and evaluating the model on each split.

3.4.3. Model Analysis and Evaluation

Model analysis and evaluation are crucial steps in building a real-world data science project. You do not want to deploy a model that has major problems or cannot generalize to new data. Assessing the performance and effectiveness of machine learning models will help determine how well the model generalizes to unseen data and whether it meets the desired objectives and determine any bottleneck that needs to be solved before deploying the model into production.

Here are key aspects to consider when analyzing and evaluating models:

- **Evaluation Metrics:** Choose appropriate evaluation metrics based on the specific task and the nature of the data. For classification tasks, metrics like accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) are commonly used. Regression tasks often employ metrics like mean squared error (MSE), root means squared error (RMSE), mean absolute error (MAE), or R-squared.
- **Confusion Matrix:** In classification tasks, analyze the confusion matrix to gain insights into the model's performance across different classes. It helps identify true positives, true negatives, false positives, and false negatives, enabling a deeper understanding of the model's strengths and weaknesses.
- **Bias and Fairness:** Assess the model for potential biases and fairness issues, particularly in sensitive applications such as loan approvals or hiring processes. Investigate if the model exhibits disparate impact or unfair treatment

towards certain demographic groups and take corrective measures if necessary.

- **Overfitting and Underfitting:** Analyze the model's performance on both the training and validation/test data. Overfitting occurs when the model performs well on the training data but poorly on the validation/test data, indicating it has memorized the training data instead of capturing underlying patterns. Underfitting occurs when the model performs poorly on both training and validation/test data, suggesting it fails to capture the underlying patterns adequately. Adjust the model's complexity or hyperparameters to mitigate these issues.
- **Feature Importance:** Assess the importance of different features or variables in the model's predictions. Techniques such as feature importance scores, permutation importance, or SHAP (SHapley Additive exPlanations) values can help identify the most influential features. Understanding feature importance can provide insights into the factors driving the model's predictions and inform future feature engineering efforts.
- **Business Impact:** Evaluate the model's performance in terms of its impact on the business or problem domain. Consider how the model's predictions can be translated into actionable insights and decision-making. Assess whether the model achieves the desired

objectives and if it aligns with the stakeholders' requirements and expectations.

- **Documentation and Reporting:** Document the analysis and evaluation process, including the chosen metrics, results, and any important findings. Prepare clear and concise reports or visualizations to communicate the model's performance and insights to stakeholders, ensuring transparency and facilitating decision-making.

3.4.4. Model Push/ Export

This step refers to the process of deploying or exporting a trained machine learning model into a production environment where it can be used to make predictions or provide insights on new, unseen data. This step is crucial for integrating the model into real-world applications and leveraging its capabilities to deliver value.

Here are the key considerations and steps involved in the model push/export process:

- **Model Serialization:** Serialize the trained model into a format that can be easily stored, transferred, and loaded into the production environment. This typically involves converting the model object into a binary representation, such as a pickle file or a serialized object in a specific format (e.g., ONNX, PMML) that is compatible with the chosen deployment framework.

- **Model Packaging:** Bundle the serialized model with any required dependencies or auxiliary files that are necessary for its execution in the production environment. This ensures that all necessary components are packaged together and can be easily deployed without missing dependencies.
- **Infrastructure Setup:** Set up the infrastructure and environment in the production system to accommodate the model's execution. This may involve configuring servers, containers, or cloud-based platforms to host and run the model efficiently and securely.

3.4.5. Inference Pipeline Implementation

Inference pipeline refers to the implementation of a systematic workflow or process that takes raw input data and generates meaningful predictions or insights using a trained machine learning model. It encompasses all the necessary steps, from data preprocessing to post-processing, to facilitate efficient and reliable inference on new, unseen data.

Here are the key components and considerations when implementing an inference pipeline:

- **Data Preprocessing**: Prepare the input data for inference by performing the necessary preprocessing steps. This may include handling missing values, scaling or normalizing features, encoding categorical variables, or applying any required transformations to ensure consistency with the training data.

- **Data Integration:** Integrate the preprocessed data with any additional relevant data sources or features that are required for the inference process. This may involve combining different data streams, merging databases, or accessing external APIs to enrich the input data.
- **Model Loading:** Load the trained machine learning model that was previously trained on historical data. This can be done by deserializing the serialized model object or loading it from a model repository or storage system. Ensure compatibility between the model and the execution environment.
- **Feature Extraction:** Extract or select the relevant features from the preprocessed data that are required for making predictions or generating insights. This step involves mapping the input data to the expected format that the model can consume.
- **Model Inference:** Apply the loaded model to the preprocessed input data to make predictions or generate outputs. This typically involves passing the input data through the model's forward pass, using the appropriate APIs or libraries provided by the chosen machine learning framework.

- **Post-processing:** Process the model's predictions or outputs to obtain meaningful results that can be easily interpreted and used for decision-making. This may involve thresholding, scaling back to the original units, converting probabilities to class labels, or performing any domain-specific transformations.
- **Result Visualization:** Visualize the results or insights obtained from the inference process in a clear and interpretable manner. This may include generating plots, and charts, or effectively communicating the predictions or insights to stakeholders, decision-makers, or end-users.
- **Integration with Production Systems:** Integrate the inference pipeline with the existing production systems or applications where the predictions or insights will be utilized. This may involve exposing the pipeline as an API, integrating it with a web application, or embedding it within an existing software system.
- **Performance Optimization:** Optimize the inference pipeline for performance and efficiency. This includes techniques such as batching or parallelizing the inference process, utilizing hardware acceleration (e.g., GPUs), or leveraging distributed computing frameworks to handle high workloads.
- **Monitoring and Maintenance:** Implement monitoring mechanisms to track the performance and stability of the inference pipeline in the production environment. Regularly monitor the

model’s accuracy, response time, resource utilization, and potential errors or issues. Perform maintenance tasks, such as periodic retraining, updating the model, or refining the preprocessing steps, as needed.

3.5. Model Deployment

Now everything is ready to deploy your machine learning into production. This step includes further steps starting from deploying the model into production, scheduling and orchestrating the training pipeline, monitoring the system and model performance, and creating alerts and notifications. Deploying an ML pipeline on the cloud for building industry-level data science projects involves leveraging cloud services and infrastructure to host and execute the pipeline components.

Here is a step-by-step guide to help you with the process:

- **Cloud Platform Selection**: Choose a cloud platform that suits your project requirements. Popular options include Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). Consider factors such as available services, pricing, scalability, and integration capabilities.
- **Data Storage**: Identify the appropriate cloud storage service to store your datasets and any intermediate results. Options include Amazon

S3, Azure Blob Storage, or Google Cloud Storage. Set up the required buckets or containers to store the data securely.

- **Compute Resources**: Provision of the necessary computational resources to run your ML pipeline components. This can be done using services such as Amazon EC2, Azure Virtual Machines, or Google Compute Engine. Choose the instance type and size based on your computational requirements.
- **Containerization**: Containerize your ML pipeline components using containerization technologies like Docker. This ensures consistent and reproducible execution across different environments. Create Docker images for each component of your pipeline, including data preprocessing, model training, and inference.
- **Container Orchestration**: Deploy and manage your containerized pipeline components using container orchestration platforms like Kubernetes or container services provided by your cloud platform (e.g., Amazon EKS, Azure Kubernetes Service, Google Kubernetes Engine). Set up the necessary clusters and configure the deployment specifications.
- **Pipeline Orchestration**: Implement pipeline orchestration using workflow management tools or services. This helps automate the execution of various pipeline stages and manage dependencies between components. Popular options include Apache Airflow, AWS Step

Functions, Azure Logic Apps, or GCP Cloud Composer.

- **Automation and Monitoring**: Implement automation and monitoring capabilities to track the performance, health, and scalability of your ML pipeline. Configure monitoring and alerting services provided by the cloud platform or integrate with third-party monitoring tools. Monitor metrics such as CPU and memory utilization, data ingestion rates, and error rates.
- **Security and Access Control**: Ensure appropriate security measures are in place to protect your data and pipeline components. Configure access control policies, encryption mechanisms, and authentication methods. Follow best practices for securing cloud resources and comply with industry-specific regulations and standards.
- **Continuous Integration and Deployment**: Implement continuous integration and deployment (CI/CD) practices to streamline the deployment process and enable efficient updates to your ML pipeline. Use CI/CD tools like Jenkins, GitLab CI/CD, or AWS Code Pipeline to automate testing, version control, and deployment of pipeline components.
- **Scalability and Cost Optimization**: Optimize your cloud deployment for scalability and cost efficiency. Configure auto-scaling policies to dynamically adjust the computaional resources based on workload demands. Use cloud services

like AWS Lambda, Azure Functions, or Google Cloud Functions for serverless execution and cost optimization.

- **Documentation and Maintenance**: Document the deployment process, infrastructure setup, and configuration details for future reference and knowledge sharing. Regularly review and update your deployment to incorporate new features, address security vulnerabilities, and optimize resource utilization.

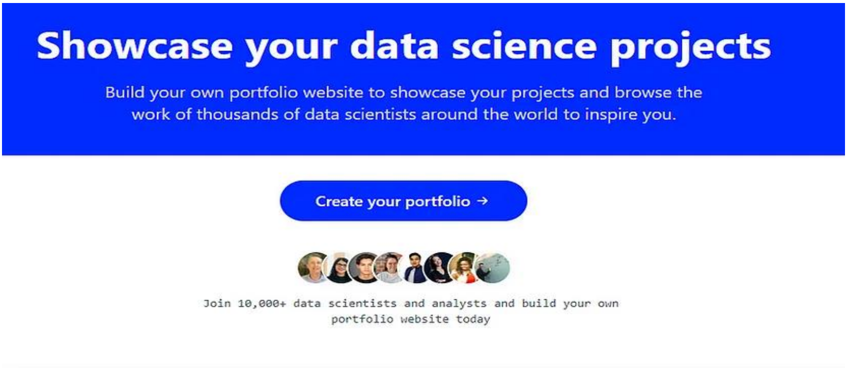
3.6. Communication & Collaboration

Usually, the final step in an industrial data science project is communicating the results and the insights you got from the data with stakeholders and the business team and collaborating with other teams to make sure that the project is delivered successfully.

For personal projects, you can communicate your results with your network by publishing your project and your insights on professional social media channels such as LinkedIn, Twitter, Medium, Kaggle, and GitHub.

To be able to represent your project professionally and in a friendly way there are a lot of tools that can be used to host your projects. Personally, I prefer to use **datascienceportfol.io.** The **datascienceportfol.io** tool will take your data science portfolio representation to the next level. You can use it to build your own portfolio website for free to showcase your projects in a recruiter-friendly way.

In addition to that, you will get a personalized URL that you can share easily; you will have all your projects in one place, regardless of where they are hosted (GitHub or Kaggle or Jupiter, etc.). Finally, you will have a beautiful design that looks neat and friendly on both desktop and mobile.



Showcase your data science projects

Build your own portfolio website to showcase your projects and browse the work of thousands of data scientists around the world to inspire you.

Join 10,000+ data scientists and analysts and build your own portfolio website today

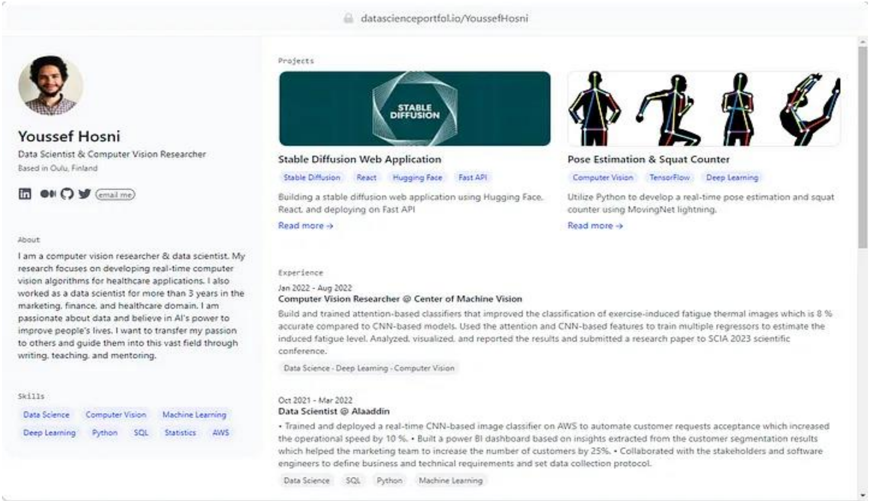
Figure 4. Showcase your data science projects using datascienceportfol.io.

Using this tool is easy:

- First, you will need to create a free account.
- Second, you can add will add your personal details and social links.
- Third, you will create pages for each project and add all the project details.

- Finally, you will add your experience and education.

Here is the final output:



datascienceportfol.io/YoussefHosni

Youssef Hosni

Data Scientist & Computer Vision Researcher

Based in Oulu, Finland

email me

About

I am a computer vision researcher & data scientist. My research focuses on developing real-time computer vision algorithms for healthcare applications. I also worked as a data scientist for more than 3 years in the marketing, finance, and healthcare domain. I am passionate about data and believe in AI's power to improve people's lives. I want to transfer my passion to others and guide them into this vast field through writing, teaching, and mentoring.

Skills

Data Science · Computer Vision · Machine Learning · Deep Learning · Python · SQL · Statistics · AWS

Projects

Stable Diffusion Web Application

Stable Diffusion · React · Hugging Face · Fast API

Building a stable diffusion web application using Hugging Face, React, and deploying on Fast API

Read more →

Pose Estimation & Squat Counter

Computer Vision · TensorFlow · Deep Learning

Utilize Python to develop a real-time pose estimation and squat counter using MovingNet lightning.

Read more →

Experience

Jan 2022 - Aug 2022

Computer Vision Researcher @ Center of Machine Vision

Build and trained attention-based classifiers that improved the classification of exercise-induced fatigue thermal images which is 8 % accurate compared to CNN-based models. Used the attention and CNN-based features to train multiple regressors to estimate the induced fatigue level. Analyzed, visualized, and reported the results and submitted a research paper to SCIA 2023 scientific conference.

Data Science - Deep Learning - Computer Vision

Oct 2021 - Mar 2022

Data Scientist @ Alaaddin

• Trained and deployed a real-time CNN-based image classifier on AWS to automate customer requests acceptance which increased the operational speed by 10 %. • Built a power BI dashboard based on insights extracted from the customer segmentation results which helped the marketing team to increase the number of customers by 25%. • Collaborated with the stakeholders and software engineers to define business and technical requirements and set data collection protocol.

Data Science · SQL · Python · Machine Learning

Figure 5. Data science portfolio was built using datascienceportfol.io.

You will also have a custom domain like this one, that you can share on your social media and resume. You can also get inspired for your next project by browsing by topic all the projects added by the community here.

4. Guided Projects: A Starting Point

Learning by doing is one of the best ways to sharpen your existing knowledge and learn and acquire new skills. This is a general rule for any field and data science is not an exception. I truly believe that one of the best ways to master the skills of the field of data science and to distinguish yourself is by doing end-to-end projects that cover different aspects of the field and it will be better to focus more on the business and companies you would like to join in the future.

Guided projects are a very important starting step on this road, they will allow you to easily start doing an end-to-end project and see the whole process without getting stuck or distracted by the many obstacles that you can meet if you work on your own from scratch. However, this should be a starting point, not the only step you take on this road. In this chapter, I will discuss more the importance of guided projects and how you can choose the best projects to do.

4.1. Learning by Doing

Learning by doing refers to a theory of education expounded by American philosopher John Dewey. It is a hands-on approach to learning, meaning students must interact with their environment to adapt and learn. This way of learning sharpens your current skills and knowledge and helps in gaining new skills that could only be acquired by doing.

Car driving is a perfect example of this, you can read as much as you would like about the theory of driving and the rules, and this is very important, and the more you understand the theory the better you get in the practical part. But you will only be able to drive better by applying this knowledge on the real road. In addition to that, there are some skills and knowledge that will only be gained by driving.

Data science is the same as driving. It is very important to have solid theoretical knowledge and to regularly increase it to be able to get better while working on a project. However, you should always apply this theoretical knowledge to projects. By this, you will deepen your understanding of these concepts and Knowledge, have a better point of view of how they work in a real-life, and will also show others that you have strong theoretical knowledge and are able to put them into practice.

4.2. Data Science Guided Projects

In the previous section, the importance of learning by doing was discussed. In this section, the idea of a data science-guided project will be further explained. Data science-guided projects are a project-based learning process in which the instructor guides you through the whole process of the project.

There are a lot of benefits to it:

- It removes the barriers between you and doing projects.
- Saves you a lot of time thinking about the project and preparing the data.
- It allows you to apply theoretical knowledge without getting distracted by obstacles.
- Practical tips that can save your effort and time in the future.
- Help you create your portfolio.
- Getting hands on skills in new technologies without spending much time and effort learning them from scratch.
- Writing professional and high-quality code.

It is important to note that the guided projects are very good as a starting point. However, it is very important that you do projects on your own using the skills you learned. The reason for that is the fact that you will meet more difficulties during self-projects which is like the one you met while working on real-world projects.

Another useful tip to make the most out of the guide projects is to always add more to the guided project after finishing it. This will help you do more unique projects and add them to your resume or portfolio and will enhance your learning experience.

5. Ten End-to-End Guided Data Science Projects to Build Your Portfolio

Data science is one of the most sought-after fields in today’s job market. With the ever-increasing amount of data being generated every day, businesses need skilled data scientists who can extract meaningful insights from the vast amount of information available. As a result, data science has become a highly competitive field, and building a strong portfolio is essential to stand out from the crowd.

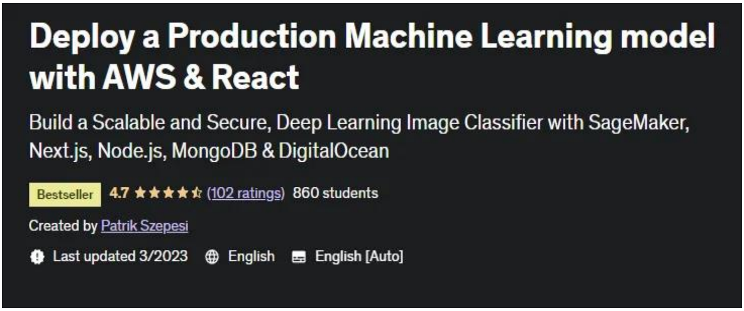
In this chapter, we have curated a list of 10 end-to-end guided projects that will help you hone your data science skills while creating a robust portfolio. These projects cover a range of topics, including data cleaning, data visualization, machine learning, and more. So, whether you’re a beginner or an experienced data scientist looking to enhance your skills, these projects will provide you with valuable hands-on experience and help you develop a well-rounded portfolio.

5.1. Automatic Speech Recognition System

The first project is building an Automatic Speech Recognition System. This is a 15-hour live implementation of an Automatic Speech Recognition System. It includes the complete project flow starting from the business problem statement to the deployment part.

5.2. Building Production-Ready Enterprise-Level Image Classifier with AWS & React

The second guided project is **Building Production-Ready Enterprise-Level Image Classifier with AWS & React** on Udemy. In this project-based course, you are going to use AWS SageMaker, AWS API Gateway, Lambda, React.js, Node.js, Express.js MongoDB, and DigitalOcean to create a secure, scalable, and robust production-ready enterprise-level image classifier.



Deploy a Production Machine Learning model with AWS & React

Build a Scalable and Secure, Deep Learning Image Classifier with SageMaker, Next.js, Node.js, MongoDB & DigitalOcean

Bestseller 4.7 ★★★★☆ (102 ratings) 860 students

Created by Patrik Szepesi

Last updated 3/2023 English English [Auto]

Figure 6. Building production-ready enterprise-level image classifier with AWS & React

You will be using best practices and setting up IAM policies to first create a secure environment in AWS. Then you will be using AWS' built-in SageMaker Studio Notebooks, where you will be shown how you can use any custom dataset you want.

You will perform Exploratory data analysis on our dataset with Matplotlib, Seaborn, Pandas, and NumPy. After getting insightful information about the dataset, you will set up the Hyperparameter Tuning Job in AWS, where you will learn how to use GPU instances to speed up training and how to use multi-GPU instance training.

You will then evaluate the training jobs and look at some metrics such as Precision, Recall, and F1 Score. Upon evaluation, you will deploy the deep learning model on AWS with the help of AWS API Gateway and Lambda functions.

You will then test our API with Postman and see if we get inference results after that is completed and will secure our endpoints and set up autoscaling to prevent latency issues. Finally, you will build our web application which will have access to the AWS API. After that, you will deploy our web application to DigitalOcean.

5.3. Predicting Data Science Salaries Application

The third guided end-to-end project is **predicting the data science salaries** by Ken Jee. In this project, you will first collect data science job requirements and expected salary data using web scraping from Glassdoor. Then the data is cleaned and explored and modeled. The model will then be put into a production environment using Flask.

Figure 7. Predicting data science salaries application

5.4. Real Estate Price Prediction Web Application

The fourth end-to-end project is a **real estate price prediction web application** by codebasics. In this guided project, you will also go through an end-to-end project to predict the real estate price. As usual, it starts with a problem statement, data collection, data cleaning, feature engineering model building, and deploying the model using Flask and on AWS EC2.

Figure 8. Real estate price prediction web application

5.5. Potato Disease Classification Mobile Application

The fifth project is building a **potato disease classification mobile applicatio**n. In this project, you will build a mobile app using React Native to classify potato disease using a deep learning model trained on the collected data and deployed on GCP.

Figure 9. Potato disease classification mobile application

5.6. Sports Celebrity Image Classification Web Application

The sixth guided project is the **Sports Celebrity Image Classification Web Application**. In this project, you will Build a website to classify sports celebrity images using a deep learning model trained on the collected data model deployed on the Flask server.

5.7. Real-Time Data Analysis Application

The seventh project is the **real-time data analysis application**. In this project, you will build a real-time data analysis application for E-commerce sales data using tools such as Kafka, Spark, Apache Cassandra, and Superset.

5.8. Machine Learning Model Monitoring using Airflow and Docker

The eighth project is **Machine Learning Model Monitoring using Airflow and Docker**. In this MLOps Project, you will learn to build an end-to-end pipeline to monitor any changes in the predictive power of the model or degradation of data.

5.9. AI-Based Hybrid Recommender System

The ninth project is building an **AI-Based Hybrid Recommender System**. This project aims to develop an AI-Based Hybrid Recommender System that combines the strengths of multiple recommendation techniques to offer more accurate and diverse recommendations to users. Specifically, the system will incorporate both content-based and collaborative filtering approaches to offer personalized recommendations based on user behavior, preferences, and similarities with other users.

The AI-Based Hybrid Recommender System will utilize machine learning algorithms and natural language processing techniques to analyze user data, including user ratings, browsing history, and product features. The system will then generate recommendations based on this analysis and provide users with a list of products or services that they are likely to be interested in.

The project will involve designing and developing the AI-Based Hybrid Recommender System, integrating it with existing systems, and testing its performance and accuracy. The project team will work collaboratively to identify the best combination of recommendation techniques and algorithms to ensure the system offers the most accurate and diverse recommendations possible.

5.10. Embedding-Based Search Engine

The last project is building an **Embedding-Based Search Engine**. This project aims to build a search engine using advanced NLP algorithms and techniques.

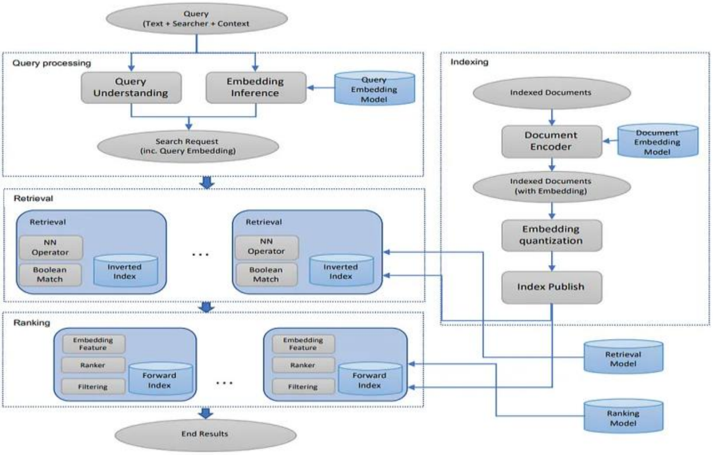


Figure 10. Embedding-Based Search Engine

6. Ten Websites with Open Datasets to Build Your Portfolio

one of the biggest challenges for aspiring data scientists is finding high-quality, unique datasets to build your portfolio with.

Fortunately, there are numerous websites that offer open datasets for free. In this chapter, we will explore 10 of the best websites to find open datasets, including well-known platforms like Kaggle and Google Dataset Search, as well as lesser-known sources like Pudding.cool and Buzzfeed.

Whether you're a beginner or an experienced data scientist, these websites are sure to provide you with the data you need to build impressive projects and advance your skills.

6.1. Google Dataset Search

Google Dataset Search is a powerful search engine that allows users to find datasets from a wide range of sources. It was launched in 2018 with the goal of making it easier for researchers, data scientists, and journalists to discover and use open datasets.
The search engine is designed to surface datasets that are hosted on a variety of websites, including academic

publishers, government agencies, and non-profit organizations. It uses a combination of metadata and structured data to index datasets and make them searchable.

One of the key benefits of Google Dataset Search is that it makes it easy to find datasets that are relevant to a specific research topic or question. Users can search for datasets using keywords, or they can use filters to narrow down their search by date, file type, or license type.

In addition to searching for datasets, users can also preview and download them directly from the search results page. The search engine also provides information on the provenance and quality of each dataset, making it easier for users to evaluate their suitability for their needs.

6.2. Kaggle

Kaggle is a popular platform that hosts a large collection of datasets and competitions for data scientists and machine learning enthusiasts. The platform was launched in 2010 and has since grown to become one of the most popular online communities for data scientists. One of the main features of Kaggle is its vast collection of datasets, which cover a wide range of topics and come from a variety of sources. Users can browse and search through the datasets on the website and download them for free. This makes it a great resource for anyone who needs to find data for their research or analysis.

In addition to its dataset collection, Kaggle also hosts a variety of machine-learning competitions. These competitions are designed to challenge data scientists to develop algorithms and models to solve complex problems. The competitions often come with cash prizes and are a great way to showcase your skills and build your portfolio.

Kaggle also provides a range of tools and resources to help data scientists get started with their projects. This includes tools for data visualization, data cleaning, and data analysis. The platform also has a large community of users who share their knowledge and expertise through forums and discussion groups.

6.3. UCI Machine Learning Repository

The UCI Machine Learning Repository is a popular resource for machine learning researchers and practitioners. It is maintained by the Center for Machine Learning and Intelligent Systems at the University of California, Irvine and provides access to a wide range of datasets that are commonly used in the machine learning community.

The datasets available on the UCI Machine Learning Repository cover a wide range of topics, including finance, healthcare, sports, and entertainment. They are typically pre-processed and formatted in a way that makes them easy to work with for machine learning tasks, such as classification, regression, and clustering. Some of the most popular datasets on the repository include the Iris dataset, which is commonly used for classification tasks, and the Wine Quality dataset, which is used for regression tasks. The repository also includes several datasets for anomaly detection, time series analysis, and text mining.

One of the benefits of using the UCI Machine Learning Repository is that the datasets are well-documented and have been used in numerous research studies and machine learning competitions. This means that there is often a wealth of resources available for working with these datasets, including published research papers, code repositories, and discussion forums.

6.4. Data.gov

Data.gov is an online repository maintained by the U.S. government that provides access to a vast collection of datasets. The website was launched in 2009 with the goal of promoting transparency and openness in government by making federal data available to the public.

The datasets available on Data.gov cover a wide range of topics, including education, healthcare, the environment, and public safety. Many of the datasets are collected and maintained by government agencies, such as the Department of Education, the Environmental Protection Agency, and the National Institutes of Health. One of the key benefits of Data.gov is the transparency it provides. The website makes it easy for citizens and researchers to access and analyze data that was previously only available to government agencies. This allows for more informed decision-making and can help to promote accountability and transparency in government.

Data.gov also provides a range of tools and resources to help users work with the datasets available on the site. This includes data visualization tools, data analysis software, and APIs that allow users to access the data programmatically.

6.5. Awesome Public Datasets GitHub Repo

Awesome Public Datasets is a curated collection of open datasets from various sources, including academic institutions, government agencies, and non-profit organizations. The collection is maintained on GitHub by the community, and it is regularly updated with new datasets.

The datasets available on Awesome Public Datasets cover a wide range of topics, including natural language processing, computer vision, healthcare, finance, and

social sciences. They are available in a variety of formats, including CSV, JSON, and XML.
One of the benefits of using Awesome Public Datasets is that the datasets have been carefully curated by experts in the field. This means that the datasets are high quality and have been used in research studies and projects. The curated nature of the collection also means that it is easier to find relevant datasets for specific research questions.

In addition to the datasets themselves, Awesome Public Datasets also provides links to tools and resources for working with the data. This includes libraries for various programming languages, data visualization tools, and tutorials for working with the datasets.

6.6. Reddit’s /r/datasets

Reddit’s /r/datasets is a community-driven forum on Reddit where users can share and discuss open datasets. The forum is a popular resource for data scientists, researchers, and students who are looking for publicly available datasets for their projects.

Users can browse through the forum to find datasets that are relevant to their research or project. The datasets on /r/datasets cover a wide range of topics, including social sciences, healthcare, finance, and natural language processing. They are often shared by individuals who have collected or compiled the data for their research or project and are willing to make it publicly available.

One of the benefits of using /r/datasets is that the community is highly engaged and active. Users can ask questions and discuss the datasets with other members of the community, which can clarify the nuances of the data and how it can be used.

In addition to sharing datasets, /r/datasets also provide links to tools and resources for working with the data. This includes libraries for various programming languages, data visualization tools, and tutorials for working with specific datasets.

6.7. Pudding.cool

Pudding.cool is a website that specializes in visual storytelling using data visualization and interactive graphics. They produce original content on a wide range of topics, including pop culture, politics, and social issues.

One of the unique features of Pudding.cool is its focus on storytelling. They use data visualization and interactive graphics to tell compelling stories and make complex data accessible to a broader audience.

In addition to its original content, Pudding.cool also provides a collection of open datasets on its website.

These datasets are often used in their own reporting and analysis and cover a wide range of topics, including music, sports, and social issues.

One of the benefits of using Pudding.cool's datasets is that they have been carefully curated by their team of data journalists and analysts. This means that the datasets are high quality and have been used in their own research and reporting.

Pudding.cool's datasets are available for download in a variety of formats, including CSV and JSON. They also provide links to relevant articles and reports that use the data, as well as resources for working with the data.

6.8. FiveThirtyEight

FiveThirtyEight is a popular website that specializes in data journalism and statistical analysis. It was founded by Nate Silver in 2008 and is now owned by ABC News. The website covers a wide range of topics, including politics, sports, economics, and culture. They are known for their unique approach to reporting, which combines data analysis with traditional reporting methods.

In addition to their articles and reports, FiveThirtyEight also provides a collection of open datasets on their website. These datasets cover a wide range of topics and are often used to support their reporting and analysis.

One of the benefits of using FiveThirtyEight's datasets is that they have been carefully curated by their team of data journalists and analysts. This means that the datasets are high quality and have been used in their own research and reporting.

FiveThirtyEight's datasets are available for download in a variety of formats, including CSV and JSON. They also provide links to relevant articles and reports that use the data, as well as resources for working with the data. Overall, FiveThirtyEight is a valuable resource for anyone who is interested in data journalism and statistical analysis. Their datasets, along with their articles and reports, provide a unique perspective on current events and trends and can be used for research and analysis in a variety of fields.

6.9. KDnuggets

KDnuggets covers a wide range of topics related to data science, including news and trends, tutorials, job listings, and educational resources. They also provide a collection of open datasets that can be used for research and analysis.

One of the benefits of using KDnuggets' datasets is that they have been carefully curated and selected for their quality and relevance to the data science community. The datasets cover a wide range of topics, including text analysis, image recognition, and natural language processing.

KDnuggets’ datasets are available for download in a variety of formats, including CSV, JSON, and SQL. They also provide links to relevant articles and tutorials that use the data, as well as resources for working with the data.

6.10. Buzzfeed

BuzzFeed is a digital media company that produces and distributes news, entertainment, and lifestyle content across a variety of platforms, including its website and social media channels.

Buzzfeed has also a lot of high-quality open datasets. It has a lot of interesting data, pre-cleaned, and with well-written commentary in the form of articles attached.

7. Five Game-Changing Free Tools to Enhance Your Data Science Portfolio

Establishing an impressive data science portfolio is integral to showcasing your talents, attracting prospective employers or collaborators, and progressing your career in data science. Compelling projects are necessary, but using the appropriate tools can take it one step further.

In this chapter, we will explore five game-changing free tools that can enhance your data science portfolio in a way that makes it more impactful and visually appealing. These tools include features such as project organization, interactive visualizations, collaborative capabilities, version control support, and reproducibility support making these additions stand out among their competition and showcase your expertise professionally and persuasively.

7.1. Datascienceportfol.io

datascienceportfol.io is a tool that will take your data science portfolio representation to the next level. You can use it to build your own portfolio website for free to showcase your projects in a recruiter-friendly way. In addition to that you will get a personalized URL that you can share easily, you will have all your projects in one place, regardless of where they are hosted (GitHub or Kaggle or Jupiter, etc.). Finally, you will have a beautiful design that looks neat and friendly on both desktop and mobile.

Using this tool is easy:

- First, you will need to create a free account.
- Second, you can add will add your personal details and social links.
- Third, you will create pages for each project and add all the project details.
- Finally, you will add your experience and education.

You will also have a custom domain same as this one "https://www.datascienceportfol.io/YoussefHosni" that you can share on your social media and resume. You can also get inspired for your next project by browsing by topic all the projects added by the community here.

7.2. Voilà

If you're a data scientist who enjoys working with Python and is interested in learning web development, then Voilà is going to be your new best friend! This amazing library allows you to effortlessly create impressive web applications and interactive dashboards using your Jupyter notebooks.

And here's the cool part: you can easily share your creations with the rest of the world! But don't just take my word for it. Look at the Voilà Gallery, where you'll find a wide range of examples that demonstrate the incredible capabilities of this library. It's the perfect way to make sure that voilà has all the features you need before you jump in and start building your own web apps.

To use Voilà you first need to install it using the following command: Voilà can be installed with the mamba or Conda package manager.

```
mamba install -c conda-forge voila
```

or from PyPI:

```
pip install voila
```

Once installed you have two options to use it:

1. As a standalone application: Voilà can be used to run, convert, and serve a Jupyter Notebook as a standalone app. This can be done via the command-line, with the following pattern:

```
voila <path-to-notebook> <options>
```

2. As a Jupyter server extension: You can also use Voilà from within a Jupyter server (e.g., after running Jupyter lab or Jupyter notebook). To use Voilà within a pre-existing Jupyter server, first, start the server, then go to the following URL:

```
<url-of-my-server>/voila
```

7.3. Dash by Plotly

Dash by Plotly is an exceptional free tool that enables you to build interactive web applications and dashboards for your data science portfolio quickly and effortlessly. Thanks to its intuitive Python framework, Dash makes creating dynamic visualizations and data-driven apps simple and effortless.

Dash is an invaluable tool for creating interactive visualizations and web applications in your data science portfolio. Leveraging its Python framework, Dash makes creating engaging visualizations simple while expanding beyond static displays with web applications, drawing upon your existing data science skillset. Customize project appearance seamlessly while using existing data science libraries seamlessly across devices ensuring responsive designs as well as real-time updates.

Through Dash, you can demonstrate your interactive data visualizations and web development expertise while offering users a compelling user experience — making your portfolio projects even more impactful and visually attractive.

7.4. DagsHub

DagsHub is an innovative platform for managing and collaborating on data science projects, offering tools such as version control, data management, and project organization to streamline workflow creation and distribution. DagsHub allows you to store all of your code, data, and models in one convenient place for easier collaboration on projects with others. In addition, DagsHub makes it easier for you to keep an eye on historical changes to code/data changes and reproduce experiments with ease.

It supports numerous data science frameworks like Jupyter Notebooks, TensorFlow, PyTorch, and Scikit-learn as well as popular tools like GitHub and Slack making integration simple into existing workflows. DagsHub provides an ideal environment for building your data science portfolio thanks to its powerful version control features, centralized organization capabilities, collaboration features, reproducibility support, data, and model management abilities, seamless integration with popular tools, and professional presentation options. Plus, its public repository gives your research greater public visibility!

DagsHub provides an effective platform to demonstrate progress, collaborate with others, ensure reproducibility, manage data and models efficiently, integrate preferred tools seamlessly, present your work professionally, and reach a wider audience — making it an essential component for building and sharing data science portfolio.

7.5. Deepnote

Deepnote is an amazing platform that allows you to host and publish Jupyter Notebooks with code and interactive outputs, enabling your projects to be showcased interactively. Deepnote makes creating collaborative data notebooks easy, whether for Python, SQL, or no-code analysis. Running securely in the cloud and connecting securely to any data source — as well as beta testing products or writing tutorials or running APIs securely from anywhere — whether for beta-testing products, writing tutorials, running APIs securely connecting databases, etc.

Starting your project off right with Deepnote is straightforward — all it requires is creating a free account, building or loading a Jupyter Notebook, then sharing it on social media or your portfolio. This platform can prove immensely helpful when conducting data analytics projects or proof-of-concept efforts.

8. Ten Portfolio Mistakes You Should Avoid

Creating a portfolio that stands out from the crowd is not always easy, and there are many common mistakes that aspiring data scientists make when putting together their portfolios.
In the last chapter of the book, we'll explore some of the most common portfolio mistakes that data scientists make and provide tips and advice on how to avoid them. From using overused or outdated datasets to not having a comprehensive README, we'll cover a range of topics that can help you to create a portfolio that truly sets you apart.

8.1. Using Overused or Outdated Datasets

Using overused or outdated datasets can be a common mistake in a data science portfolio. While it's perfectly acceptable to use well-known datasets like the iris dataset or the Titanic dataset to learn and apply new skills. However, relying on these datasets can make your portfolio seem unoriginal or outdated. This can be a problem, especially if you are looking for employment opportunities, as potential employers may be looking for candidates who can apply their skills to new and challenging problems.

One way to avoid this mistake is to use a variety of datasets in your portfolio, including newer and less well-known datasets. For example, you might consider using datasets related to emerging technologies or industries, such as cryptocurrency, genomics, or climate science. This can show potential employers that you are keeping up with current trends and can apply your skills to new and innovative areas.

Another approach is to focus on datasets that are relevant to specific industries or domains, such as healthcare, finance, or transportation especially if you have a domain of interest that you are trying to work in. By using datasets that are specific to these industries, you can demonstrate your ability to work with real-world data and solve practical problems that are relevant to specific business contexts.

Ultimately, the key is to use datasets that demonstrate your skills and expertise in a way that is relevant and compelling to potential employers or clients. By using a variety of datasets, including newer and less well-known ones, you can showcase your versatility and adaptability, and position yourself as a data scientist who can tackle a wide range of data-related challenges.

8.2. Having Only Online Courses Guided Projects

Having only online courses guided projects in your data science portfolio can be a mistake, as it may not accurately reflect your ability to work on real-world data science problems independently. While online courses and guided projects can be a great way to learn new skills and gain experience with different tools and techniques, they are often designed to be relatively straightforward and may not fully capture the complexity and messiness of real-world data science projects.

To avoid this mistake, it is important to include a mix of projects in your portfolio that demonstrates your ability to work independently and tackle real-world data science problems. This might include projects that you have completed as part of internships, research projects, or personal side projects.

When selecting projects for your portfolio, it’s important to focus on those that highlight your problem-solving skills, ability to work with large and messy datasets, and ability to apply advanced modeling techniques to extract insights from data. You should also try to showcase your ability to effectively communicate your findings and insights to both technical and non-technical stakeholders.

By including a mix of projects in your portfolio that demonstrates your ability to work independently on real-world data science problems, you can showcase your versatility and ability to deliver value in a variety of settings. This can be particularly appealing to potential employers or clients who are looking for data scientists who can hit the ground running and deliver results quickly and efficiently.

8.3. Not Having End-to-End Projects

Not having end-to-end projects in your data science portfolio can be a major mistake, as it can prevent potential employers or clients from understanding your problem-solving skills in a real-world context. End-to-end projects refer to projects that involve the entire data science workflow, from data collection and cleaning to modeling and deployment.

By including end-to-end projects in your portfolio, you can demonstrate your ability to work with messy, real-world data and develop practical solutions that can be implemented in a production environment. This can be especially valuable for potential employers or clients who are looking for data scientists who can help them solve complex problems and drive business value.

Without end-to-end projects in your portfolio, you may be missing out on opportunities to showcase your skills and expertise in a way that resonates with potential employers or clients.

8.4. Lack of Project Diversity

Diversity is an important aspect of any data science portfolio, as it demonstrates your versatility and ability to solve a range of problems using different techniques and technologies. By showcasing a variety of data science projects in your portfolio, you can demonstrate your adaptability and problem-solving skills and show potential employers or clients that you can apply your skills in a range of contexts.

When selecting projects to include in your portfolio, it's important to consider a range of factors, such as the types of data analyzed, the techniques and tools used, and the industries or domains in which the projects were completed. For example, you might include projects that involve machine learning, data visualization, natural language processing, or predictive modeling, to showcase your proficiency in different areas.

Another way to demonstrate diversity in your portfolio is to include projects that touch on a range of industries or domains, such as healthcare, finance, retail, or education. This can show potential employers or clients that you have a broad perspective and can apply your skills in a range of contexts.

8.5. Failing to Showcase Your Domain Expertise

Failing to showcase your domain expertise can be a mistake in your data science portfolio, especially if you have experience or expertise in a specific industry or field. Many data science projects involve working with data from specific domains, such as healthcare, finance, or marketing, and having knowledge and experience in these areas can be an asset.

To avoid this mistake, it’s important to highlight your domain expertise in your portfolio and demonstrate how it has informed your approach to data science. For example, you might include projects that involve analyzing healthcare data to identify patterns and trends in patient outcomes or developing predictive models for financial markets based on economic indicators.

In addition to showcasing your domain expertise through specific projects, you might also consider including blog posts, articles, or other content that demonstrates your understanding of key issues and challenges in your field. This can help to position you as a thought leader in your industry and can be particularly appealing to potential employers or clients who are looking for data scientists with specific domain knowledge.

8.6. Not Having a Comprehensive Readme for Your Projects

It is important to have a comprehensive documentation readme file for each project to be able to show your work in a proper way, or else you will not be able to show up your work and results.

A comprehensive project readme should have the following sections, according to Andrew Jones:

- **Project Overview:** First, you should provide the project background and motivation and what is the goal of the project.
- **Data Overview & Preparation Steps:** A brief overview of the data used, how you collected the data, and how you prepared it, which includes data preprocessing and feature engineering. Also, it is also important to provide a summary of the data.
- **Methodology Overview:** You should also mention the techniques and tools used to achieve the project goals.
- **Results & Outcomes**: Finally, you should also mention the results and the project outcomes and what they mean.
- **Growth & Next Steps:** You should also mention what are the next steps and what you would do if you had more time.

8.7. Lack of Code Comments & Documentation

Lack of code comments and documentation can be a common mistake in data science portfolios. While it’s important to showcase your technical skills and expertise, it’s equally important to provide clear and concise explanations of your code and methodology, so that others can understand and replicate your work.

To avoid this mistake, it’s important to include detailed comments and documentation in your code, explaining how it works and why you made certain decisions. This can include comments that explain the purpose of each line of code, as well as longer explanations of your methodology and approach. This is important as it will show that you can work efficiently on a production-level code and can fit in with a team.

In addition to code comments, you might also consider including a README file or other documentation that provides an overview of your project, including the problem you were trying to solve, the data you used, and the techniques you applied. This can be particularly helpful for others who may want to build on your work or replicate your results.

8.8. Not Having Well Structured Projects & Codes

When potential employers or clients review your portfolio, they are looking for evidence that you can develop well-organized, modular, and maintainable code, as well as demonstrate the ability to apply best practices in software engineering.

To avoid this mistake, it's important to structure your projects and code in a way that is easy to follow and maintain. This includes breaking your code into reusable functions and modules, using appropriate naming conventions, and organizing your files and directories in a logical way.

A very good starting point is the project structure shown in the figure below, which was provided by Abhishek Thakur in his book Approaching (Almost) Any Machine Learning Problem and was shown in figure 1.

8.9. Lack of Reproducibility

Reproducibility is an essential aspect of data science, as it allows others to understand and verify your work. By providing detailed instructions on how to reproduce your work, including the software, packages, and libraries used, you can ensure that others can replicate your methodology and findings.

In addition to that it will be easier for others to understand and build upon your work, reproducibility also helps to ensure that your analysis is transparent and trustworthy. If your work cannot be reproduced, it can be difficult for others to trust your findings or use them to make informed decisions.

To ensure reproducibility in your portfolio, it's important to document your work in a clear and organized manner, including any assumptions or decisions made along the way. You should also provide clear instructions on how to install and use any software, packages, or libraries used in your analysis, and include code snippets or scripts that can be easily run.

In some cases, it may also be useful to provide access to the data used in your analysis, either by including it in your portfolio or by providing a link to a public repository where it can be accessed.

By prioritizing reproducibility in your data science portfolio, you can demonstrate your commitment to transparency and accuracy, and make it easier for potential employers or clients to understand and evaluate your work.

8.10. Not Publishing Your Projects on social media

Not publishing your data science projects on social media can be a mistake, as it can limit your exposure to potential employers or clients who may be interested in your work. Social media platforms like LinkedIn, Twitter, and GitHub are popular among data scientists and can be a great way to share your work, build your brand, and connect with others in the industry.

By sharing your projects on social media, you can showcase your skills and expertise to a wider audience, including potential employers or clients who may be looking for data scientists with your specific skill set. This can be particularly important if you are just starting out in the industry and don’t have a large network or reputation yet.

In addition to showcasing your work, social media can also be a great way to get feedback and suggestions from other data scientists. By sharing your work on platforms like GitHub or Kaggle, you can invite others to review and critique your code and methodology and learn from their feedback and suggestions.

Publishing your data science projects on social media can help you build your brand, showcase your skills, and connect with others in the industry. By doing so, you can increase your visibility and attract more opportunities, whether that be job offers, consulting projects, or collaborations with other data scientists.

Afterword

Thanks for purchasing and reading my book!
If you have any questions, feedback or praise, you can reach me at: Youssef.Hosni95@outlook.com

I would be happy if you connect with me personally on LinkedIn:
https://www.linkedin.com/in/youssef-hosni-b2960b135/

If you liked my writings, make sure to follow me on Medium:
https://medium.com/@youssefraafat57
Subscribe to my newsletter to never miss any of my writings:
https://youssefh.substack.com/

Youssef Hosni is currently a research scientist at VTT. Before that, he worked in different domains as a researcher and data scientist for four years. On the side, Youssef has helped thousands of data scientists improve through his writings.

Subscribe to my newsletter for more career tips:
https://youssefh.substack.com/