





EXPLORE || DIGITAL SKILLS

Regression Predict Instructions

Individual Predict - Tailored for Part-time Students

Regression Predict Summary

In this predict you will be tasked to build and deploy a ML model and to participate in a Kaggle challenge.

Who	<div>You</div> <div></div>	You are tasked with building a model for the government of Spain to predict the shortfall between energy generated by fossil fuels and energy generated with renewable sources.
What	<div>Regression Models</div> <div></div>	Supervised machine learning techniques, such as the ones covered throughout this digital skill, will be used to build a model to forecast the target values.
How	<div>Python</div> <div></div>	You are free to use any relevant regression and/or time series method(s).
Why	<div>Learn</div> <div></div>	The purpose of this predict is to guide you through the typical steps of a real-world data science projects from initial EDA, to model development and deployment and finally to communication of results.

At the end of this predict you'll have a ML prototype in production to show to recruiters, friends, and family

What is expected from you?

Task 1: Data Analysis and Model Creation

Task 2: Kaggle Challenge

Task 3: Model Deployment (bonus)

Task 4: Results Discussion

Conclusion



Mission Breakdown

In this predict you will be exposed to an **end-to-end data science problem**, which has **four distinct** tasks.

What is expected from you?

1

Notebook analysis



- As part of the first step you are required to: explore the provided dataset, clean the data, engineer new features, and build an accurate regression model.

2

Kaggle challenge



- Next you need to use your trained ML model to participate in a Kaggle competition.

3

Model deployment



- In the third step you will again use your trained ML model, this time to deploy it on a Flask webserver with the help of AWS. This is a bonus step, it is not required in an individual predict, but it is great exposure to using APIs.

4

Communication



- Finally you will be required to present your modeling process, findings and results to a group of technical and non-technical stakeholders.
- Two options: You can set up a session in person, or you can record yourself, and upload the recording / recording link to ATHENA.

What is expected from you?

Task 1: Data Analysis and Model Creation

Task 2: Kaggle Challenge

Task 3: Model Deployment (bonus)

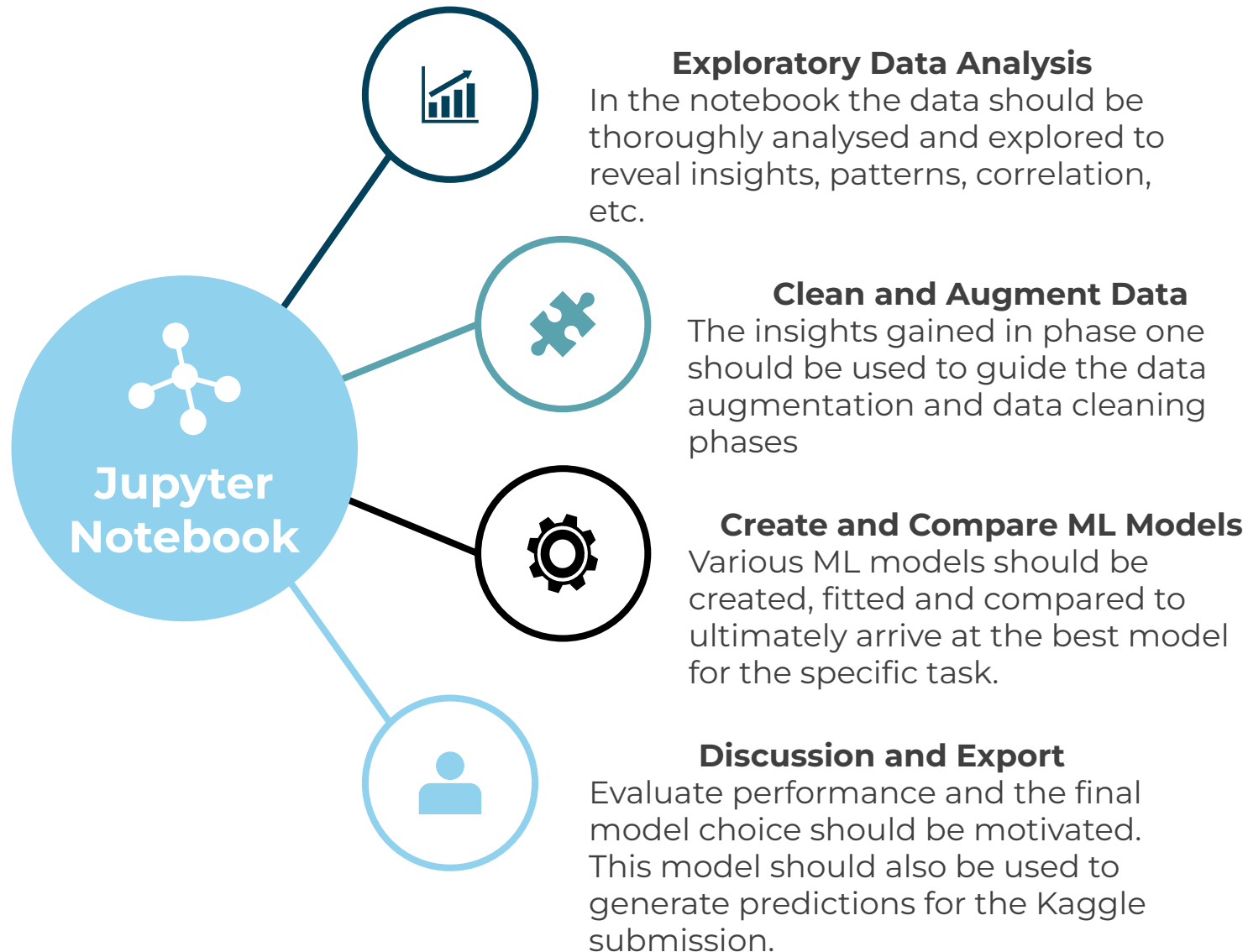
Task 4: Results Discussion

Conclusion



Developing your Model in Jupyter Notebook/Lab

The **Data Analysis and Model Creation** phase can be broken down into 4 separate sub-tasks, each with their own outcomes.



Outcomes
<ul style="list-style-type: none"> Outcome 1: EDA. Data analysed, and thoroughly understood. Outcome 2: Preprocessing. Data cleaned and relevant features added, or existing features augmented. Outcome 3: Modelling. Various ML models trained and optimised. Outcome 4: Validation. Best model thoroughly discussed and used to generate Kaggle submission.

What is expected from you?

Task 1: Data Analysis and Model Creation

Task 2: Kaggle Challenge

Task 3: Model Deployment (bonus)

Task 4: Results Discussion

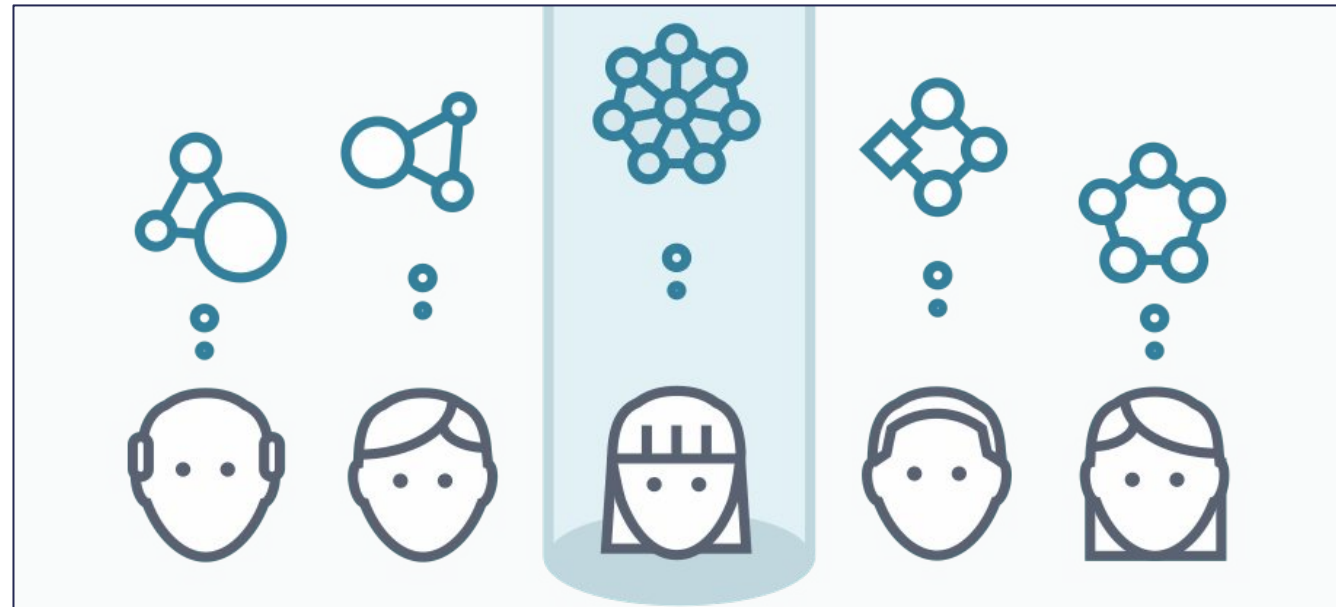
Conclusion



Compete in a Hackathon Challenge on Kaggle

In this task, you will be introduced to **Kaggle**, a fantastic online platform for doing data science projects.

Kaggle, according to their website, allows users to: discover and publish their own data sets, explore and build models in a web-based data-science environment; work with other data scientists; and enter competitions to solve data science challenges.



Your challenge for this task will be to compete in your first Kaggle competition by training a regression model to predict three-hourly demand shortfall.

How to Get Started

First things first, sign up to Kaggle and navigate to the **Spain Demand Shortfall Regression Challenge**.

1. [Sign up to Kaggle](#) and **create your own personal profile**.
2. Enter the [Kaggle Challenge](#).
3. **Use your notebook and ML model created in Task 1.** At least 1 submission is required per team.
4. **Ensure your notebook can produce a valid submission.** Submit this output to Kaggle to be placed on the Challenge Leaderboard (you can do this multiple times).
5. It is recommended to host this notebook on GitHub, and forward the repo link to your supervisor.



Competition Rules

It is important to note that this is an **Individual** challenge.

1. This project is an **individual** project. You are advised to collaborate in ideas, discussions, and direction. However, you are **not allowed to share your code, solutions, or submissions** with others.
2. **Submissions on Kaggle must be done individually** - part of your Predict **mark will be based upon your best score on the Leaderboard**.
3. You will be **required to prove how you obtained a given submission result**, which will be done through the notebook upload on ATHENA, which must include the full code used to get from the raw data, to the output submission on Kaggle. Failure to be able to prove how the kaggle submission was attained, will receive a mark of 0 for the predict.



What is expected from you?

Task 1: Data Analysis and Model Creation

Task 2: Kaggle Challenge

Task 3: Model Deployment (bonus)

Task 4: Results Discussion

Conclusion

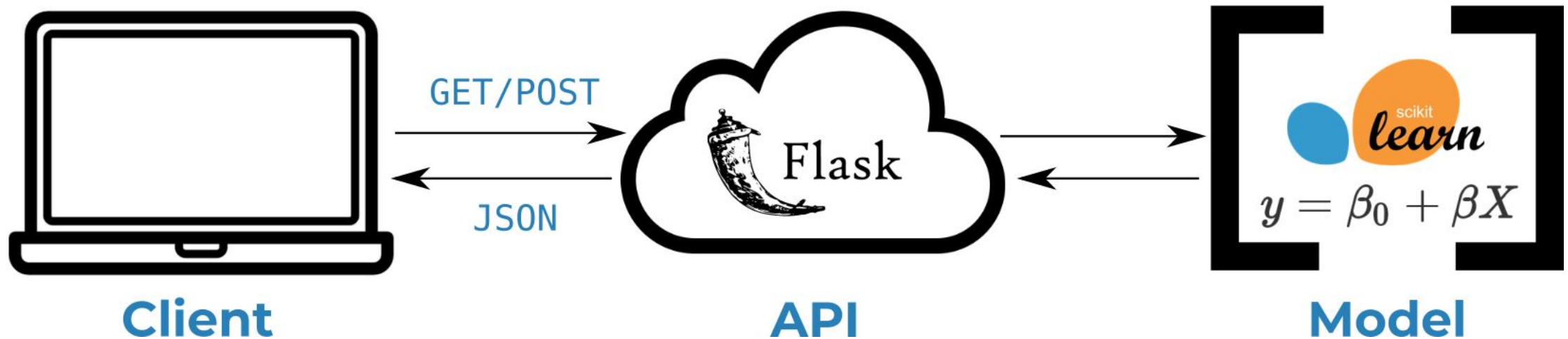


Building a Web-based API for your Model (optional)



An Application Programming Interface, or API, provides a **standardised way for other individuals to interact** with a system, program or **model you've built**.

Within the Regression Predict, **we will help you build your own API using the Flask web server framework**. You will then go on to host this API within an AWS EC2 instance, making your model available to your friends, workplace, future clients, and essentially the world at large!





Instructions to Deploy your Regression Model (optional)

Deploying your ML model ensures that it is live and on the web, for you to show to recruiters, friends and family.

1. We have created a [template repo](#) on GitHub to help you setup your API.
2. You need to **Fork the repo***.
3. Send the URL of your new GitHub repo to your supervisor.
4. While the template provides much of the scaffolding needed to run your API, you must **modify the code** to enable the API to **serve your model developed in Task 1** of the predict.
5. The repo contains instructions on how to setup the API on a local computer, as well as on an EC2 instance. Please follow these instructions carefully.
6. **We will test** that your **API** is functioning **as soon as the submission closes**. Marks will be given if an API provides a valid POST response containing a prediction from your own model.

* Forking a repo creates an exact copy of it. The copied repo, or fork, exists separately from the original i.e. changes to the fork don't affect the original repo.



What is expected from you?

Task 1: Data Analysis and Model Creation

Task 2: Kaggle Challenge

Task 3: Model Deployment (bonus)

Task 4: Results Discussion

Conclusion



Communicate your Findings: Documenting your process

In data science, communicating your findings, solutions, and results are just as important, if not more, than the actual building of the solution.

As a Data Scientist, you need to **continually communicate** your work and findings to various audiences. Within this Predict, you will be required to **explain your work to fellow data scientists**.

You will do this in the following ways (1. **Code usability**, and 2. **Presenting your findings**):

1. **Code usability.** As you **develop your solution notebook for Task 1**, you will need to **ensure that your work is fully documented and is reproducible** by a technical individual. To do this, it needs to:
 - **Have logical structure**, with an introduction, body and conclusion.
 - **Contain essential steps** within your model development process, such as an Exploratory Data Analysis (EDA), data preprocessing, modelling, performance evaluation, and model analysis.
 - Be supported with **appropriate visuals and metrics**.
 - Separate these sections and **explain your work using Markdown** cells.
 - Contain well written code which is sufficiently commented and **meets best coding practices**.

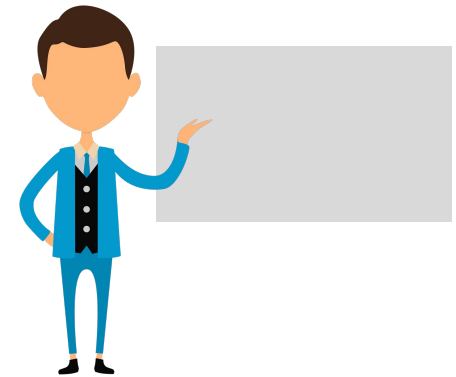


Communicate your Findings: Virtual presentation

To show the EDSA facilitation team what you have accomplished during the predict, you are required to present back to your supervisor and classmates.

2. **Present your findings** back to your supervisor and classmates (in person, or via recording).

- **Using a slide deck, provide an overview** of the salient stages in your modelling pipeline developed in Task 1 to a technical audience. Your presentation should **explain your approach, design choices, and findings** throughout the entire process.
- **Make extensive use of graphs and visuals** to provide a data-driven argument for your solution.
- Presentations to last **5-7 minutes**, and we will allow an additional few minutes of question time. Recommended structure, across a recommended 4-8 slides:
 - i. Problem statement (What are we doing? Executive summary of the objective.)
 - ii. Exploratory data analysis (Dive into relevant, insightful thoughts and visuals.)
 - iii. Approach (What did you do and how well did it work?)
 - iv. Insights (What did you find out/learn?)
 - v. Conclusion (How well did you solve the problem? What could you do to improve this?)



Presenting your work can be done live... or via recording!

You can record yourself using any software you prefer, and share a recording with your supervisor.

3. If not in person | you may alternatively record yourself presenting

- Several options are available for you to record your findings
- Teams / [OBS](#) / [YouTube Studio](#) / [Loom](#)
- Check out the links, each of the platforms provide comprehensive “how-to” introductions and should be relatively intuitive to work through

4. Tips for presenting virtually!

- Lights, camera, action! Turn your **camera on**, and **illuminate yourself**!
- Make sure we can see your **head & shoulders**
- Keep the **background noise** to a minimum
- Give us the best chance of seeing & hearing you clearly!



Drag and drop a file you want to upload
Your video will be private until you publish it

SELECT FILE

What is expected from you?

Task 1: Data Analysis and Model Creation

Task 2: Kaggle Challenge

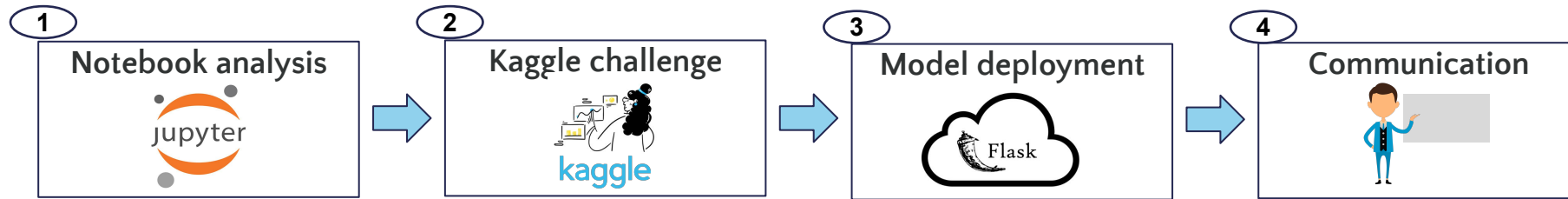
Task 3: Model Deployment (bonus)

Task 4: Results Discussion

Conclusion



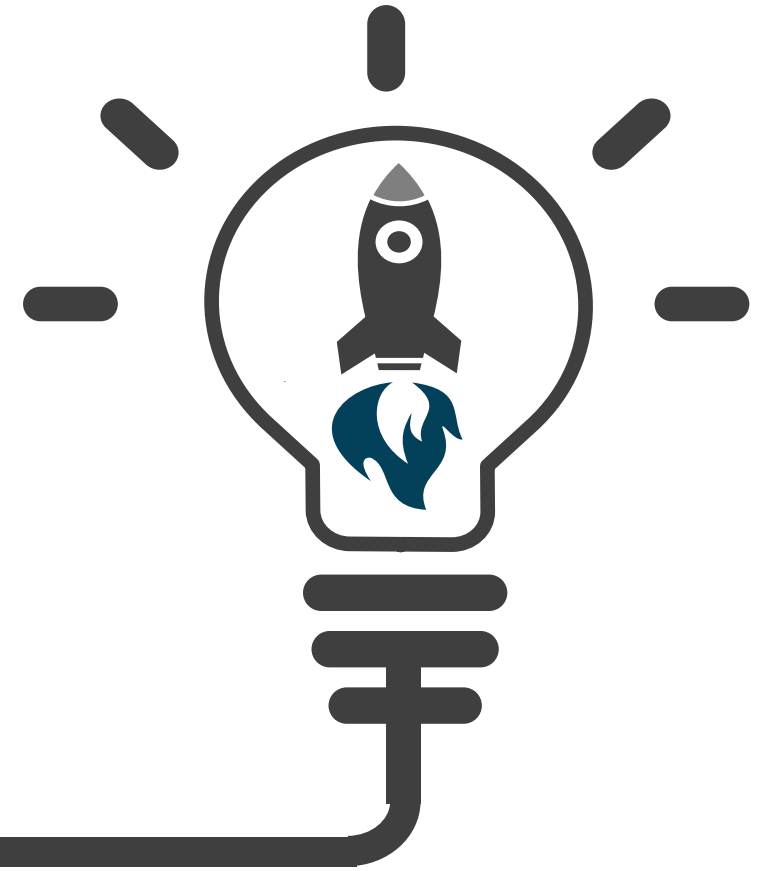
Conclusion



Remember to check the FAQ section for common questions which may arise around the predict. If you detect any problems/bugs, please create an issue and we will do our best to resolve it as quickly as possible.

Please ensure that you thoroughly analyse your data before diving in to fitting multiple ML models. The trick to succeed in this predict lie in both the feature engineering and the model building.

Good luck and have fun!!



Rubric

Presentation		35
Presentation Structure		5
Framing	Problem Statement	
	Presentation Flow	
Presentation Skills		25
Visualisation	Use of Visuals	
	Intuitive and Appealing Deck	
Speaking	Articulation of Key Concepts	
	Engaging and Confidant Speaking	
Complimentative	Deck + Talking Cohesive	
"Above and Beyond"		5
Exceptional Presentation	Discretionary	
Notebook		45
Kernel Usability		20
Functionality	Data Analysis	
	Predictive Ability	
	Cohesive Pipeline (Loading to Submission)	
	Data Science Application	
Kernel Readability		20
Report style notebook	Coherent Story	
	Use of Comments	
	Use of Markdown	
	Good Coding Practices	
"Above and Beyond"		5
Exceptional Notebook	Discretionary	
Leaderboard		20
Kaggle Score	Graded on RMSE between x and z	
Bonus Model API and Solution Delivery		15
Functional API	Using Flask to Deploy the Solution	
Delivery on AWS	Using EC2 to host the	
Version Control	Correct use of GitHub	
Project Management	PM Artefacts e.g. Trello	
Marks Available		115
TOTAL		100

Communication



Notebook analysis



Kaggle challenge

Model deployment

