Sea Turtle Conservation AI Project

Aayesha Islam

Abstract

This report contains the process and results of training CogUniversity's AI Chatbot Cammy. The training exercises focused on three aspects: answering questions about the biology of sea turtles, predicting the weight of sea turtles based on given parameters of length, width, and species, and image identification of different sea turtles species, namely Loggerhead, Green, Hawksbill, and Ridley. Throughout the process, data analysis, comprehension of the software, testing, and experimentation was needed. This report details the accuracy of Cammy's results and their practical implementation.

Training an AI Chatbot to Promote Education on Sea Turtles

Objective

Sea Turtles are an important part of the Ocean's biodiversity and conservation is very important. Cammy is trained in order to be the pinnacle of public education about sea turtles, by taking data as input and outputting crucial information in a manner easy to digest.

Assumptions

Due to the limitations of Cammy's models, certain assumptions are made regarding the input data. Firstly, for basic biology responses, it is assumed that questions will be posed in a manner understandable to English speakers and utilize somewhat proper grammar. Additionally, there is an assumption that users possess a foundational knowledge of sea turtles, including common traits like egg-laying, as questions are framed with this understanding. Secondly, in the context of weight prediction, it is presumed that the provided data is accurate, having been sourced externally. The model further assumes that each sea turtle species exhibits a weight range determined by its length and width. Thirdly, for image-to-species predictions, the assumption is that users will submit images of mature sea turtles underwater, given the nature of all input data. Consequently, Cammy

may encounter challenges when analyzing images of sea turtles on beaches or in any other surrounding area.

Input Data

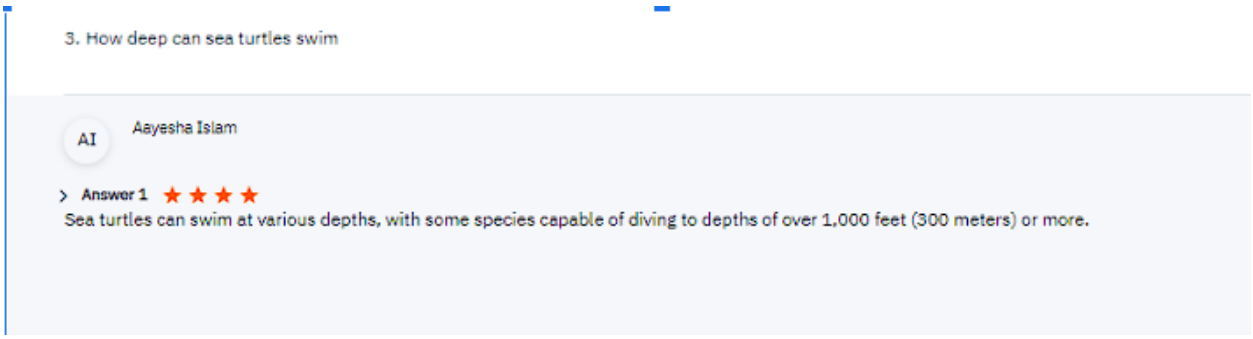
The data used to train Cammy varies according to the category, as illustrated below.

Basic Questions

Cammy was fed questions from a CSV file. There are varied questions about their species, size, lifestyle, eggs, and how to observe them. The following is a screenshot of the CSV file with the entries of the questions.

What if see a turtle laying its eggs	STB1002
How deep can a sea turtle go	STB1002
how many eggs do turtles lay at a time	STB1002
What to turtles eat	STB1002
what time of year do sea turtles lay eggs	STB1002
What guides a baby sea turtle to the water	STB1002
How deep can sea turtles dive?	STB1002
What should I do if I sea a sea turtle laying eggs on the beach	STB1002
What is the sea turtles favorite food	STB1002
How big was the largest sea turtle ever recorded?	STB1002
Can I watch the moma sea turtle lay eggs or will that disturb them	STB1002
what is the world record sea turtle	STB1002
What to sea turtles eat	STB1002
Can sea turtles swim down deep	STB1002
How old do sea turtles get	STB1002
What do sea turtles like to eat	STB1002
How long to turtles live	STB1002
When do sea turtles nest?	STB1002
When do turtles nest.	STB1002
How can I watch a sea turtle lay eggs without bothering them	STB1002
How does the freshly hatched sea turtle who which way to walk to get to water	STB1002
How many eggs are in a turtle nest	STB1002
How big can a sea turtle be	STB1002
How do sea turtles locate the ocean while they are on the beach?	STB1002
how big do sea turtles get	STB1002
what is their nesting season	STB1002
how many sea turtle eggs do you find in a single nest	STB1002
How long do sea turtles live?	STB1002
How many eggs do sea turtles lay?	STB1002
Do sea turtles live a long time	STB1002
What is the max depth a sea turtle can go	STB1002
How deep can sea turtles swim	STB1002

Cammy is also provided answers to some of the questions and provided a 1 to 4 star rating for the answers it outputs. The following is a screenshot of a sample answer.

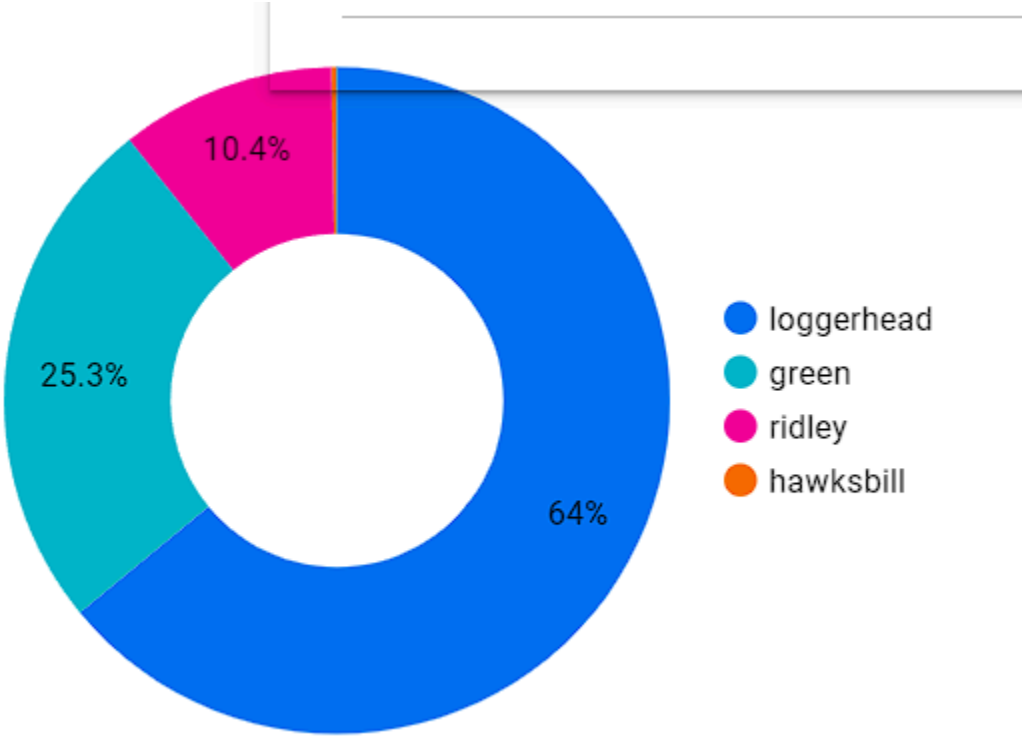


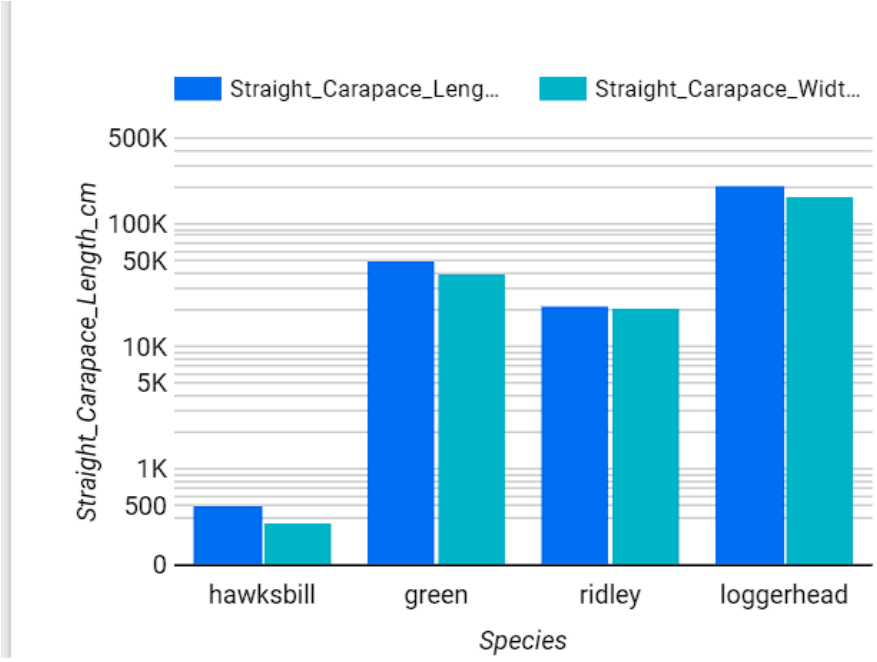
Weight Prediction

The data provided to Cammy was in a CSV file which had 5699 values, covering different datapoints such as Species, Carapace length and width, and weight.

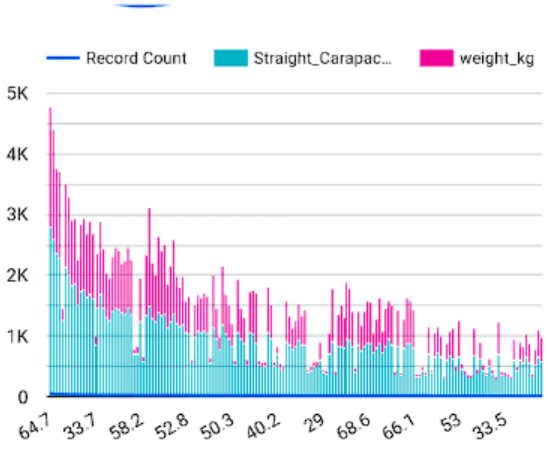
	A	B	C	D
1	**Species**	**Straight_Carapace_Length_cm**	**Straight_Carapace_Width_cm**	**weight_kg**
5679	green	19.7	15.6	0.8
5680	green	27.1	21.7	2.4
5681	loggerhead	64	50.5	37.8
5682	green	31.4	24.8	3.8
5683	loggerhead	73.4	63.6	70.1
5684	green	32.1	24.2	3.8
5685	ridley	40.1	37.7	8.2
5686	loggerhead	60.3	49.1	30
5687	green	32.8	26	4.4
5688	loggerhead	53	46.3	53.9
5689	ridley	40.6	39.7	9.6
5690	loggerhead	64.4	52.5	38.4
5691	loggerhead	65	53.2	34.2
5692	green	32.5	26.7	4.2
5693	loggerhead	64	54.6	40.8
5694	green	31.6	26.6	2.8
5695	green	30.1	26	3.6
5696	green	42	34.3	8.3
5697	loggerhead	64.3	53.2	35.8
5698	loggerhead	60	50.4	32.8
5699	loggerhead	59.3	49.7	31

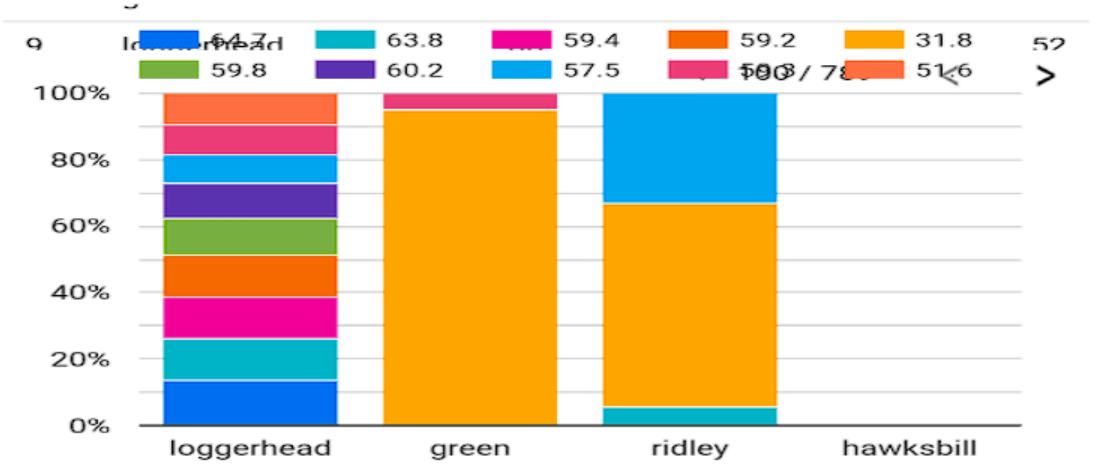
The following charts are based on the CSV file and show details about the dataset, including trends, proportions, and averages,





	Species	weight_kg	Record Count
1.	green	2.8	70
2.	loggerhead	31	68
3.	green	4	66
4.	loggerhead	37	56
5.	green	3.6	53
6.	green	3.8	53
7.	green	4.2	52

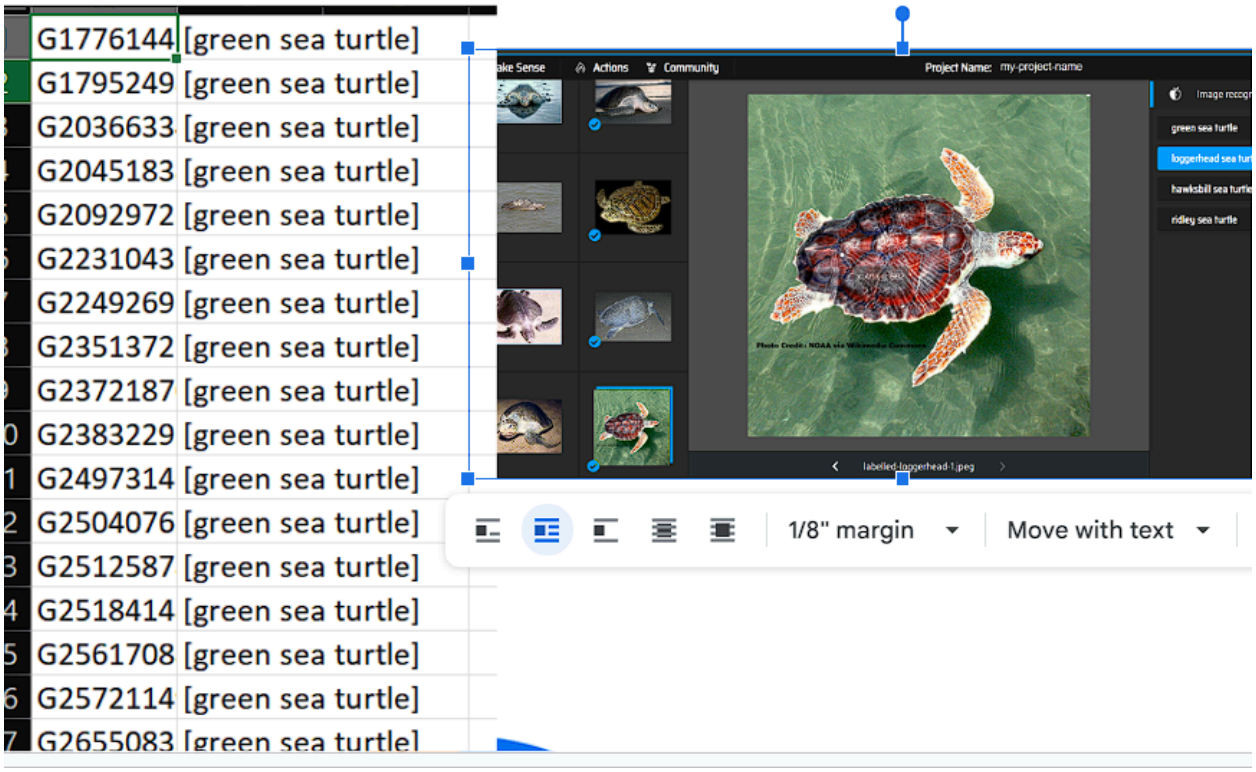




Analyzing the data notifies that there are 4 species: Loggerhead, Green, Ridley, and Hawksbill. The bar chart shows that usually turtles have proportionate length and width of the Carapace. Since turtle shells are so unique and interesting, this sort of information is crucial.

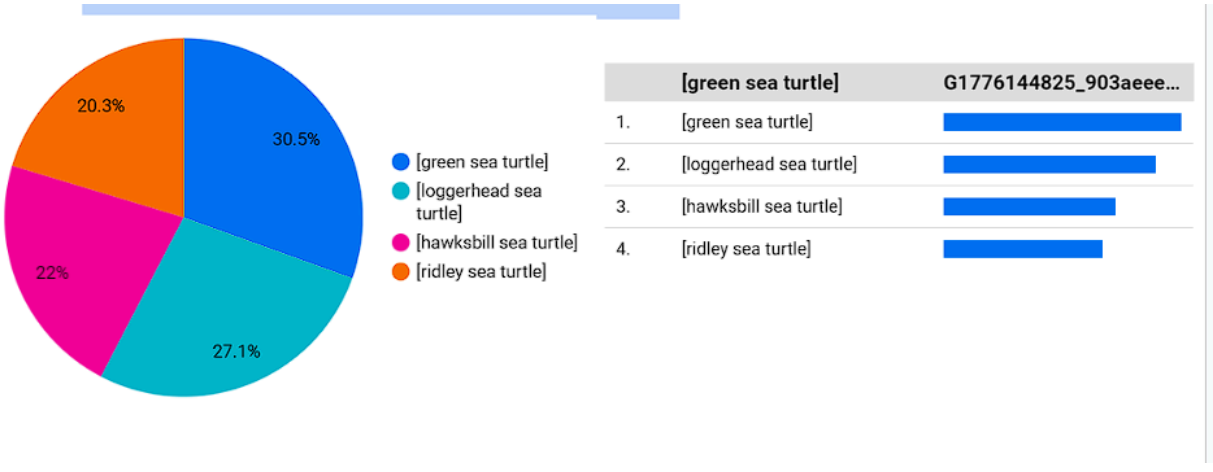
Image Analysis

The data provided to Cammy were 59 images which were stored, labeled, and converted into a CSV file using the makesense website. The CSV has the links to each image and the label given by the user.



The above two images show the Data labeling process and the resulting CSV file.

The graphs below show the distribution of species, showing there were not equal samples of data given for each species, which may cause inaccurate results.



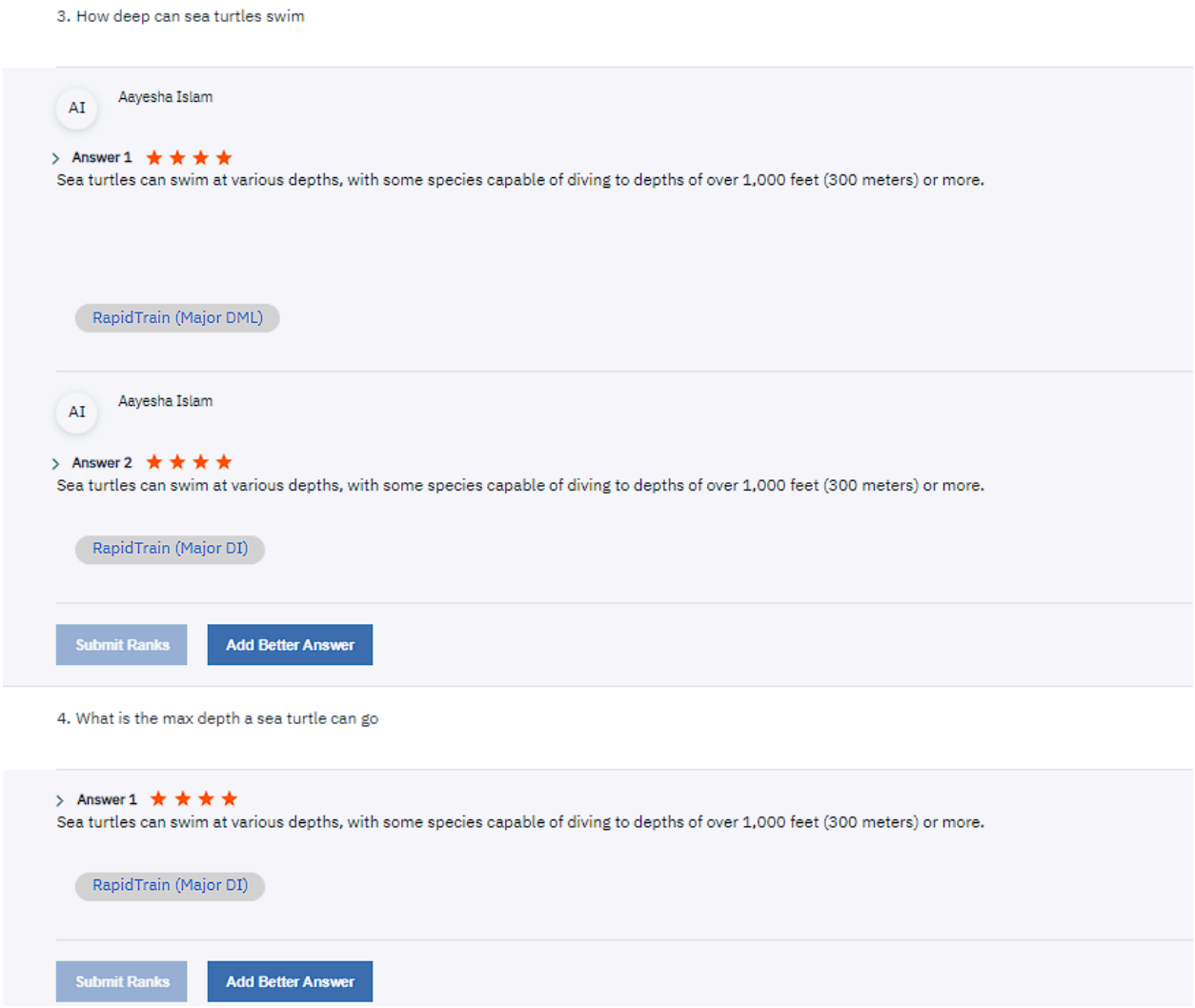
Model Conceptualization

After analyzing the results, it can be concluded that Cammy uses classification and regression as models to predict answers and provide accurate output. Classification and regression are two types of supervised machine learning algorithms. Classification deals with assigning data to predefined categories, while regression focuses on predicting numerical values based on input features. Both the basic biology and image analysis functions utilize classification to identify the most suitable answer to a question or accurately determine the species depicted in an image. The bot employs manual classifications and leverages its knowledge gained from the initial training dataset to predict the outcome for new images or questions. In the context of the basic questions, this includes analyzing a given question and classifying it into a specific "type" with a consistent answer. In the case of images, the process is simpler, involving the classification of new images based on the provided image data into the correct species. The weight prediction function, on the other hand, employs regression using the training height and weight data available in the original dataset. Regression, a fundamental statistical concept, enables the bot to mathematically predict the likely weight of a described sea turtle with ease.

Model Presentation

To comprehend and present the models used by Cammy, it is necessary to understand the concepts with simple explanations. An example which best describes classification in this context is how your email is sorted into “spam” and “not spam”. A way this is done is scanning and detecting certain content such as offers for discounted products or services (e.g., "Get 50% off on luxury watches!"), requests for personal information or financial details (e.g., "Claim your prize by providing your bank details"), phishing attempts pretending to be official communication (e.g., "Verify your account to avoid suspension"), and suspicious links or attachments. If this is found, those emails are classified as a certain “type” and thus put into the spam folder, the rest being put into the inbox. This algorithm can comprehend similar phrases and sentences of a certain “type” and understands what behavior it should output.

For the basic questions, this is how Cammy classifies words and phrases into “types” and can detect what the optimum answer would be for a question which is similar to a question where the answer was provided by the user. It also uses the rating system to correct itself and learn new information.

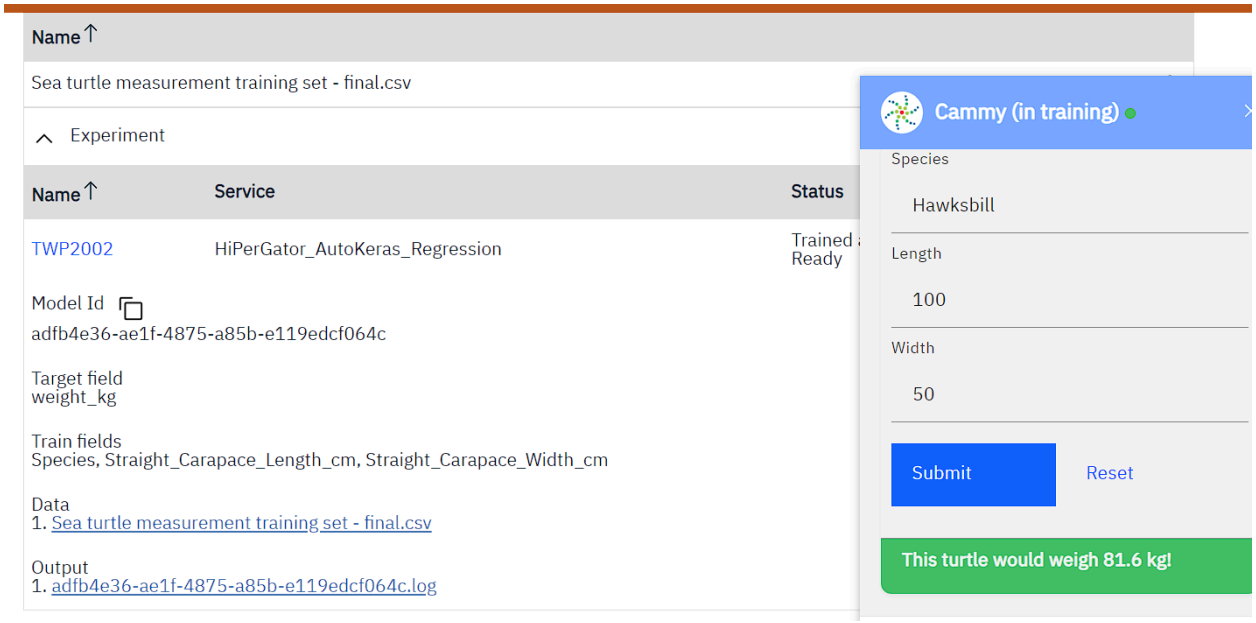


Each unique question needs an answer and a period of waiting time for Cammy to process the answers and learn from them. This is the part where Cammy is trained. Initially, Cammy's answers are not good and either unrelated or incorrect. The user ranks Cammy's answers as low, and then adds a better answer, ranking the new answer highly. Halfway through the question list, Cammy's answers start to get on topic but they are still not fully answered. To get the perfect answer for a common question, an

answer can be added from the Corpus answers. Gradually, Cammy's answers start to get better, as it gets training based on the ranking it is given for the answers. It can be seen in the following answers how Cammy automatically gives the correct answer based on the training it got from the first question. A similar classification model is used for the Image analysis aspect of Cammy, as each image is classified as a specific species based on common features.

For the weight prediction, a regression model is used. Regression is a type of mathematical modeling used in machine learning to predict numerical values based on input data. Imagine a set of data points representing houses, where each data point includes information like the house's size, number of bedrooms, and location. Regression helps build a model that can understand the relationships between these features and predict a continuous outcome, such as the house's price. The goal is to create a formula or pattern that, when applied to new data (e.g., a house with specific features), provides an estimate of the expected numerical result (e.g., the predicted price). So, in simpler terms, regression helps us make educated guesses about numerical outcomes by identifying patterns in the data we have. Similar is the case with sea turtles, as Cammy analyzes the different patterns and trends in the data parameters. The linear regression

model allows it to predict the weight based on previous analysis of given data.



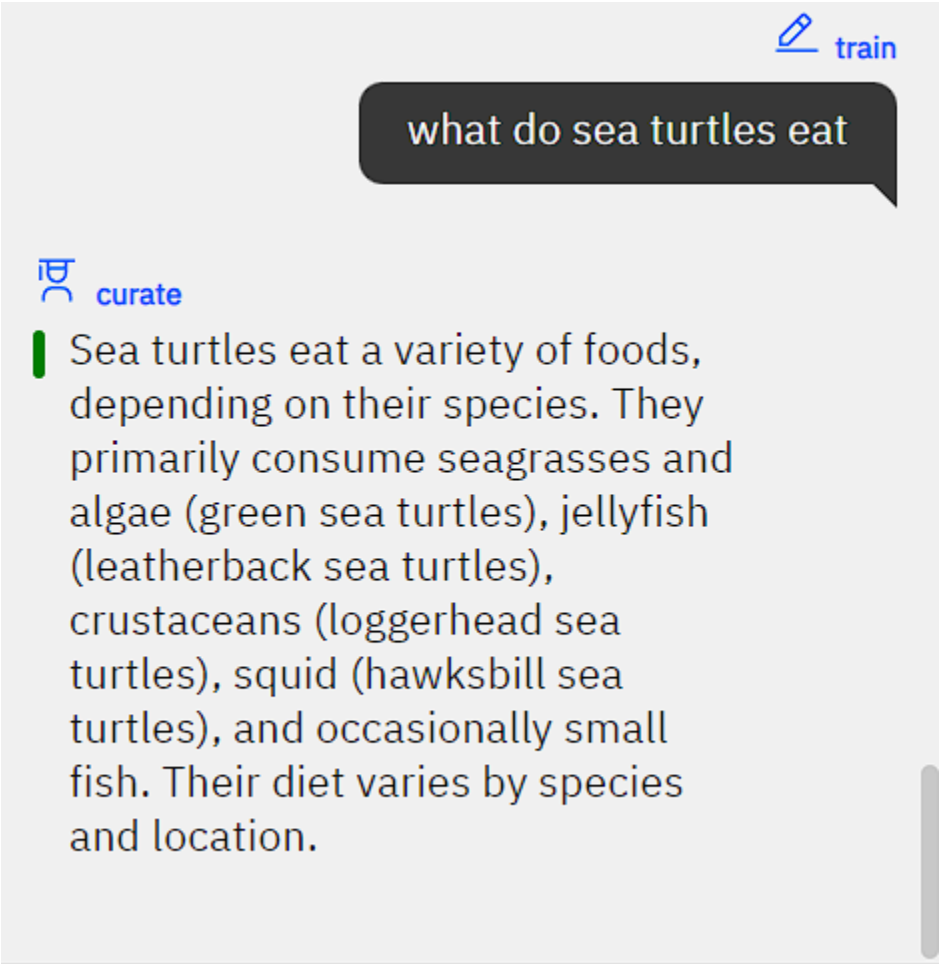
In the above screenshot, the results of Cammy's training can be seen which are displayed through an interactive chatbox, making it user friendly. Cammy does her predictions by detecting trends from the data, and the more data is given to Cammy, the more accurate the results are.

Verification and Experimentation

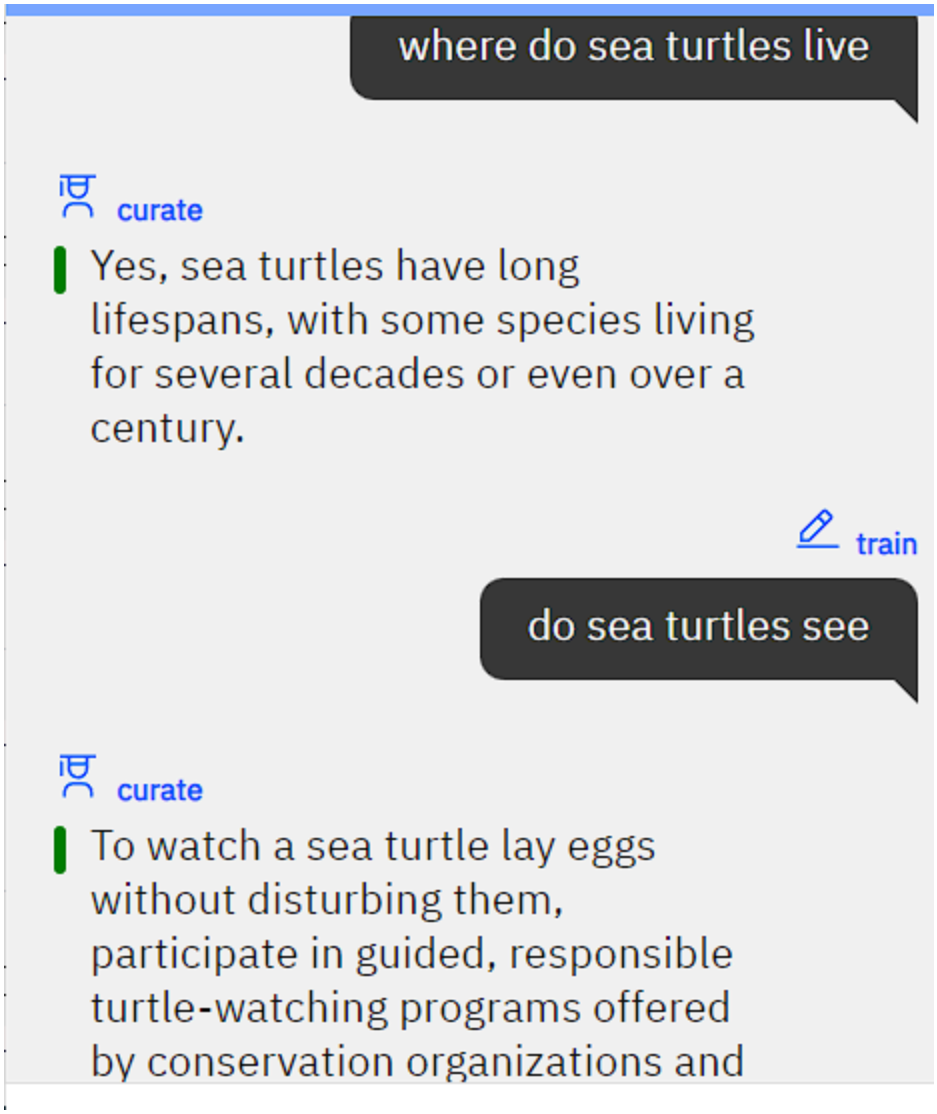
To verify the accuracy of Cammy, some edge cases were tested and experimentation was done with asking questions and inputting values which were not the norm.

Basic questions

After training, Cammy has a decent skill of answering questions correctly, as shown in the below screenshot.



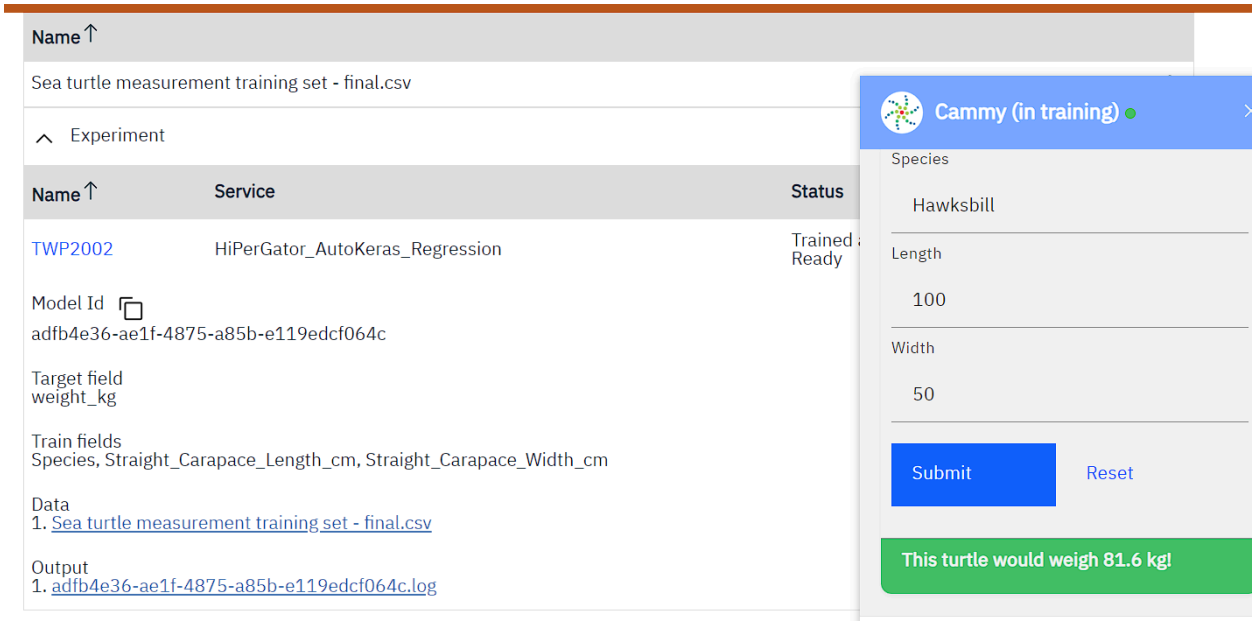
However, if the questions are tweaked in any way or something unexpected is asked, Cammy makes mistakes, as shown below.



This shows that Cammy does not have enough knowledge to form as many classifications as needed. This shows the limits of the data and that it can be improved.

Weight Prediction

Cammy is able to predict the weight of a turtle given the species, Carapace length, and width.



In the above, the results of Cammy's training can be seen which are displayed through an interactive chatbox, making it user friendly. Cammy does her predictions by detecting trends from the data, and the more data is given to Cammy, the more accurate the results are.

Cammy still does not account for edge cases or if the user puts in invalid data. The weight prediction gets stuck at 0.1 kg as shown below, showing that Cammy needs more fine tuning and data to become truly accurate.

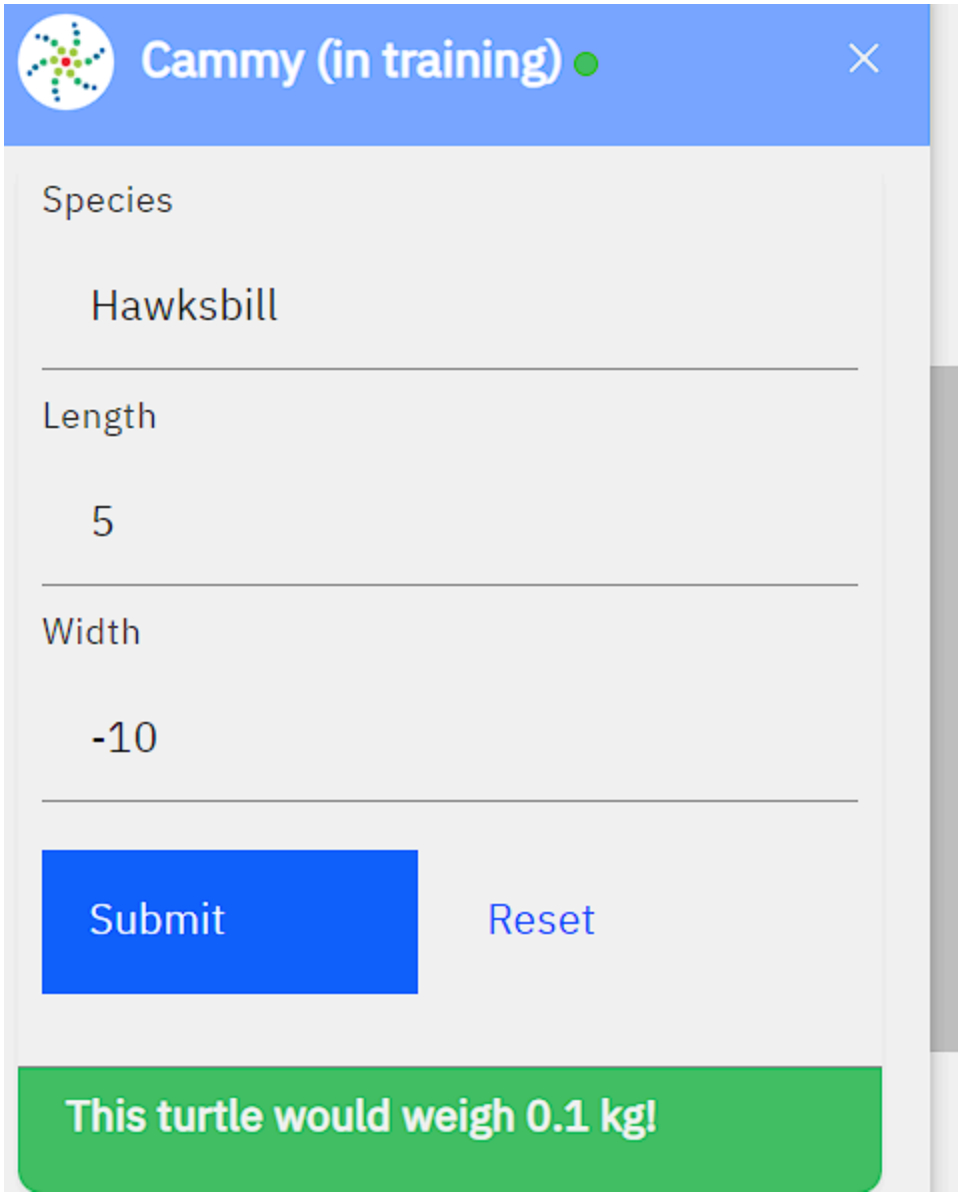
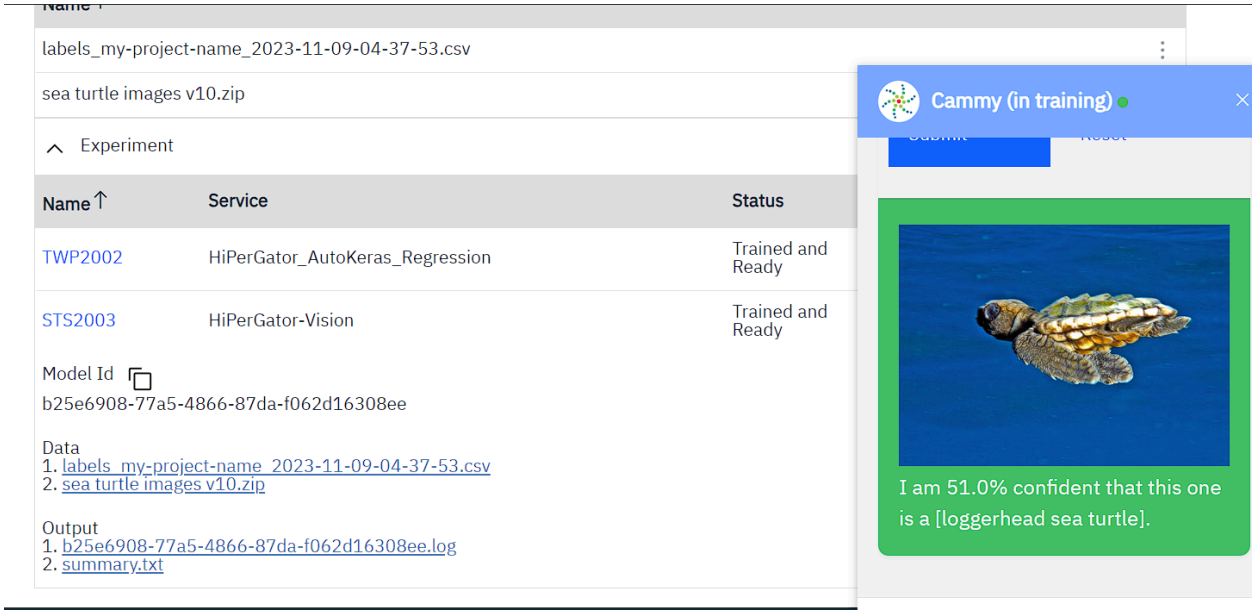


Image Analysis

Cammy performs with confidence when tested with new images, but it depends on the species, as on some species it has more data and some it has less. In the example below, it can guess Loggerhead with 51% confidence, which is a value that definitely should be improved.



Conclusion

Cammy can help people and tourists from the local and international community to learn more about turtle conservation so that they are more aware of their impact. Working with Cammy taught me how many different types of data can be collected and combined to reach important conclusions and predictions. We need predictions often in the real world, such as weather, natural disasters, job outlooks, and medical operation success chances. These various industries have been collecting data for a long time, and that effort can be brought to life using machine learning and proper modeling.

My hope is that these models will be further improved by collecting more diverse and representative data to train the models. Additionally, optimizing model hyperparameters, such as the learning rate and regularization, can lead to better predictive accuracy. The amount of data needed for a machine learning regression analysis can vary based on the complexity of the problem and the model used.

For image recognition, it is possible for Cammy to make mistakes because a lot of sea turtles have very similar features and the pictures are taken from different angles and environments that make it confusing, especially as the color palette is very similar for all the pictures. If the images were to be changed, it would be better to use high quality images of objects with more distinct features.

References

"CogUniversity," cu2.uf-cap1.cogability.net. https://cu2.uf-cap1.cogability.net/#/en-us/ (accessed Nov. 26, 2023).

Appendices

No appendices were needed in this report.