PROJECT MILESTONE 3 REPORT

PROJECT NAME	ChatBot Flight Booking System	**DATE OF STATUS ENTRY**	November 19, 2023
PROJECT CODE	BDM3035 T3F23 CFBS	**PERIOD COVERED**	November 06, 2023 to November 19, 2023
PROJECT MANAGER	Auradee Castro	**DATE OF COMPLETION**	December 15, 2023
MEMBERS	Auradee Castro (c0866821), Bhumika Babu (c0867081), Miraj Sinya (c0863371), Olivia Deguit (c0878491), Roger Mais (c0863147)		

PROJECT STATUS

PROJECT STATUS	ON TRACK	**SUMMARY**	Completed tasks for searching and canceling booked flight tickets and integrating a real-time flight tracking page in the chatbot. Conducted exploratory data analysis for ML development by applying basic pre-processing techniques before testing the dataset with pre-trained models.
TASKS	Successfully finished the tasks involving the search and cancellation of booked flight tickets in the flight booking system, and the integration of real-time flight tracking page in the chatbot. Additional validations and minor enhancements will be conducted as next steps.		
	Completed the exploratory data analysis for ML development by applying basic pre-processing techniques prior to testing the datasets using pre-trained models		

PROJECT COMPONENTS

COMPONENT	STATUS	OWNER / TEAM	NOTES
BUDGET	UNDER	MetaMorph Team	Utilize DynamoDB as an exclusive database for simplified data pipeline instead of creating a new one in AWS specifically for batch processing
RESOURCES	ON TRACK	MetaMorph Team	Tools installation and configuration completed - Python and Jupyter Notebook - Power BI Desktop - AWS account and services - Landbot - Github Repository - Microsoft Teams
TIMELINE	ON TRACK	MetaMorph Team	Project is progressing as planned without major blockers
SCOPE	ON TRACK	MetaMorph Team	Pre-trained models for sentiment analysis and passport reader instead of training and building models from scratch

PROJECT PLAN ADJUSTMENTS: SCOPE AND DESIGN REVISIONS

TASK NUMBER	DESCRIPTION	OWNER / TEAM	NOTES
CFBS-03	Simplified AWS data pipeline by using single database for both streaming and batch processing	MetaMorph Team	Utilize DynamoDB as the database for streamlined management and a simplified data pipeline, consolidating the entire process into a single database instead of creating a new data pipeline in AWS specifically for batch processing. Benefits: 1. Cost Efficiency: reduced development and AWS resource expenses 2. Simplified Maintenance: easier troubleshooting, updates and overall maintenance

CFBS-12	Pre-trained models for sentimen analysis and passport reader models	MetaMorph Team	Pre-trained models will be used instead to get the sentiment on customer reviews and to extract text information from passports, removing the complexity of developing models from ground up. 1. A minimum of two pre-trained models will be utilized, and their accuracy will be compared to decide for the final models: a) RoBERTa and VADER for sentiment analysis, and b) PaddleOCR and Pytesseract for text extraction on US passports 2. Changed dataset for sentiment analysis, which will be used to evaluate and compare the accuracy of RoBERTa and Vader models. Here's the new dataset: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

WORK ACCOMPLISHED

TASK NUMBER	DESCRIPTION	OWNER / TEAM	NOTES
CFBS-07	Flight booking feature (subtask)	Roger Mais	Successfully completed the tasks related to searching for and canceling booked flight tickets in the feature development of flight booking system. See video on Booking search and cancellation: https://drive.google.com/file/d/1Y93CUtjYqsCrc8FY4TT9b-1U1olff2hj/view?usp=drive_link
CFBS-08	Real-time Flight Tracking feature (subtask)	Roger Mais	Successfully completed the tasks on integrating real-time flight booking page into chatbot app. See video on https://drive.google.com/file/d/1Se2puLmReY9b8mwx-Ozhp5z5mA7dQVYh/view?usp=sharing
CFBS-12	Exploratory data analysis for sentiment analysis	Auradee Castro, Bhumika Babu, Olivia Deguit	Conducted data wrangling and data analysis to understand and get preliminary insight on the dataset. Since pre-trained model will be utilized for sentiment analysis, only fundamental data pre-processing techniques are applied to cleanse the reviews: - Check for missing values and duplicate records - Removing non-grammatical text like URLs, email address, hashtags - Removal of emojis - Handling slang words - Removal of extra whitespaces and stopwords Link of the codes: https://github.com/abccastro/ChatBot-Online-Flight-Booking/blob/main/Flight%20Sentiment%20Analysis.ipynb
CFBS-12	Exploratory data analysis for passport reader	Auradee Castro, Miraj Sinja	Same from previous milestone: Employed various sets of image preprocessing techniques on US passport): - Applied grayscale and noise reduction to the images - Implemented thresholding on the images Link of the codes: https://github.com/abccastro/ChatBot-Online-Flight-Booking/blob/main/Passport%20Reader.ipynb

RISKS AND ROADBLOCKS

TASK NUMBER	DESCRIPTION	OWNER / TEAM	FIX / RESOLUTION
CFBS-13	Complexity of developing ML models from scratch, potentially resulting in less accurate outcomes due to the constraints on limited dataset during the model-building process	MetaMorph Team	Utilized the pre-trained models a) RoBERTa and VADER for sentiment analysis, and b) PaddleOCR and Pytesseract for text extraction on US passports

CFBS-14	No direct data connection between Power BI and AWS DynamoDB	MetaMorph Team	The team will be utilizing the third-party application, Simba DynamoDB ODBC Drive, to test and establish connection between Power BI and AWS DynamoDB

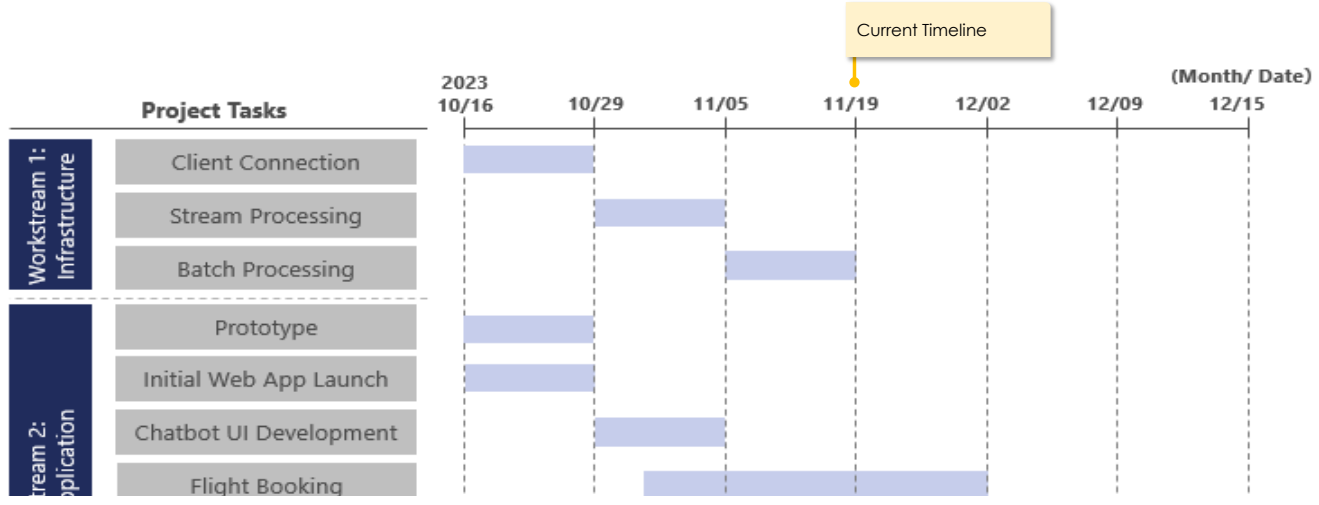
HIGHLIGHTS AND KEY TAKEAWAYS

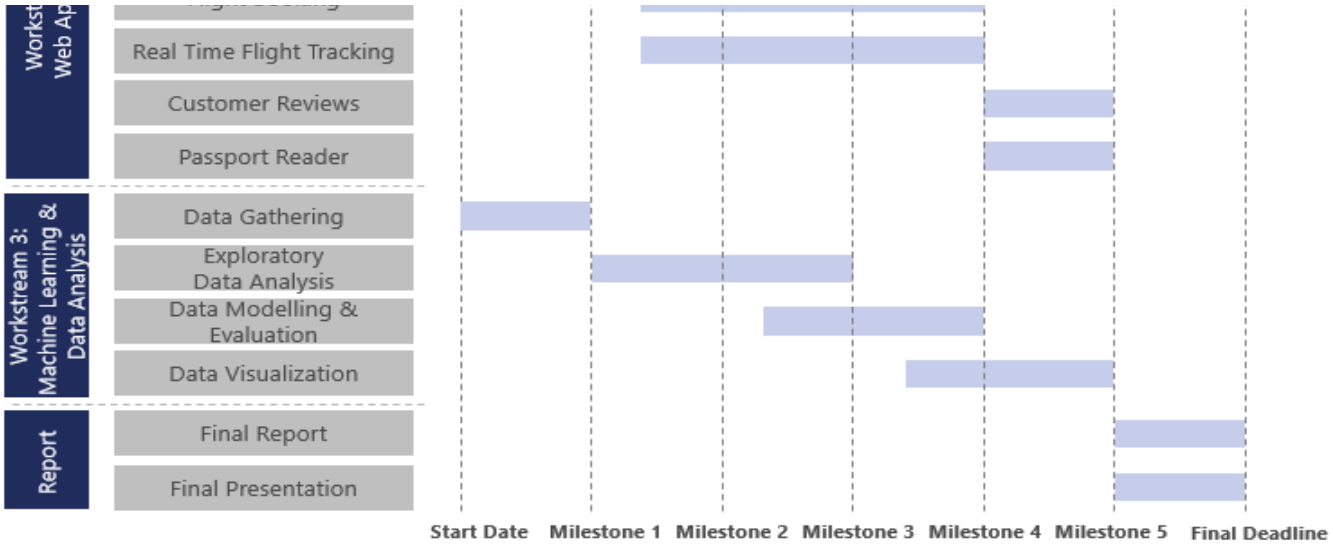
Fulfilled all requirements outlined for milestone 3 as communicated to stakeholders, except for tasks related to batch processing services. Rather than developing a new suite of AWS services for batch processing, the team chose to utilize DynamoDB as the database for streamlined management and a simplified data pipeline, consolidating the process into a single database for cost and maintenance efficiency
Pre-trained models, including RoBERTa and VADER for sentiment analysis and PaddleOCR and Pytesseract for text extraction from US passports, will be employed and evaluation to determine the final models, mitigating the complexity and potential accuracy issues associated with developing machine learning models from scratch with a limited dataset
Held a stakeholder meeting to collect feedback on modifications to the project design and tasks concerning machine learning models. Ensured thorough communication of the changes and obtained the necessary approvals.

UPCOMING WORK

TASK NUMBER	STATUS	DETAILS
CFBS-07	ON TRACK	Fully functional flight booking feature, encompassing ticket booking, searching, and cancellation, along with the integration of extra validations on input fields.
CFBS-08	ON TRACK	Fully functional real-time flight tracker with simulated FlightAPI's Flight Tracker API
CFBS-13	ON TRACK	Finalized the selection of the pre-trained model for analyzing sentiment in flight reviews by comparing and evaluating the accuracy of results from both RoBERTa and VADER models.
CFBS-13	ON TRACK	Finalized the selection of the pre-trained model for extracting text from US passports by comparing and evaluating the accuracy of results from both PaddleOCR and Pytesseract
CFBS-14	ON TRACK	Initiate the data visualization process by establishing a connection between Power BI and AWS DynamoDB. Then conduct initial analysis on the dataset essential for making informed business decisions. (December 26)

PROJECT MILESTONE AND TIMELINE




Workst
Web Ap
Real Time Flight Tracking
Customer Reviews
Passport Reader
Workstream 3: Machine Learning & Data Analysis
Data Gathering
Exploratory Data Analysis
Data Modelling & Evaluation
Data Visualization
Report
Final Report
Final Presentation
Start Date
Milestone 1
Milestone 2
Milestone 3
Milestone 4
Milestone 5
Final Deadline