StaffAny

Artificial Intelligence/Machine Learning (AI/ML) Pipeline for Sales Foundation Model

Loo Wee Sing, Caspian Michael
A0214109R
CA Report
School of Computing
Faculty of Computer Science

Introduction

This Final Year Project (FYP) aims to enhance the sales process of StaffAny, an HRMS Solutions Company, through the development of an AI for the Sales Team. The main problem faced is a complex sales workflow that distracts the sales team from essential sales tasks, ultimately contributing to a reduced ability to close more clients and generate more revenue.

To address this, an AI-driven AI/ML pipeline is built by leveraging data from StaffAny's CRM tools and communication platforms as training data. The resultant models trained from this data are then used to streamline sales operations by targeting 3 key components of the existing workflow - the **sales event**, the **sales funnel** and the **sales cycle**.

Early tests of the pipeline show promising signs of simplifying the sales team workflow and improving the sales team focus on **increasing the number of sales events that can be executed**.

Problem

Despite using CRM tools like HubSpot, communication technologies from Twilio, Aircall and Apollo.io and productivity toolchains provided by ChatGPT+, the sales team still faces performance plateaus.

Some common issues include a protracted **sales cycle** while closing clients (up to a year or more), difficulty in transitioning individuals through the **sales funnel** through qualifying individuals as potential clients and obstacles while executing **sales events** like choosing the right delivery message effectively.

Many of these issues stem from an over fragmented context when sales personnel attempt to comprehend data across different data repositories whilst preparing for sales events. This is caused by insufficient integration among the various platforms used by the sales team.

This problem is further worsened as StaffAny is a growth stage startup, with limited resources available to centralize data collection, requiring costly operational overhauls and tooling upgrades - an endeavor StaffAny is unable to support. This supports a strong thesis for AI to unify insights across these discrete data repositories, substantially enhancing the sales team ability to generate revenue.

Objectives

The resultant trained models and new AI applications from the AI/ML pipeline are judged on their ability to influence the following metrics,

Metric	Description	Formula
Total Appointments Set	Total number of appointments scheduled with potential leads.	Count of appointments
Appointment Setting Rate	Percentage of cold contacts that result in an appointment.	(Appointments set / Cold contacts) * 100
Total Number of Sales Activities	Total count of sales-related actions taken (calls made, emails sent, etc.).	Count of sales activities
Conversion Rate	Percentage of appointments that result in a sale.	(Sales closed / Appointments set) * 100
Sales Cycle Duration	The average time taken from initial contact to closing a sale.	(Sum of all sales cycles) / (Number of sales closed)

Scope

The FYP aims to meet the objectives by analyzing historical data from HubSpot and any other possible data source on sales events to build the AI/ML pipeline, train the early models and subsequently, build any further AI application layers that can fit into the existing sales operations.

This pipeline is built to be extensible, with the ability to accommodate more training data and feedback loops to enhance model learning. This helps to train more models to generate further insights into sales operations and fine-tune the trained models from the AI/ML pipeline.

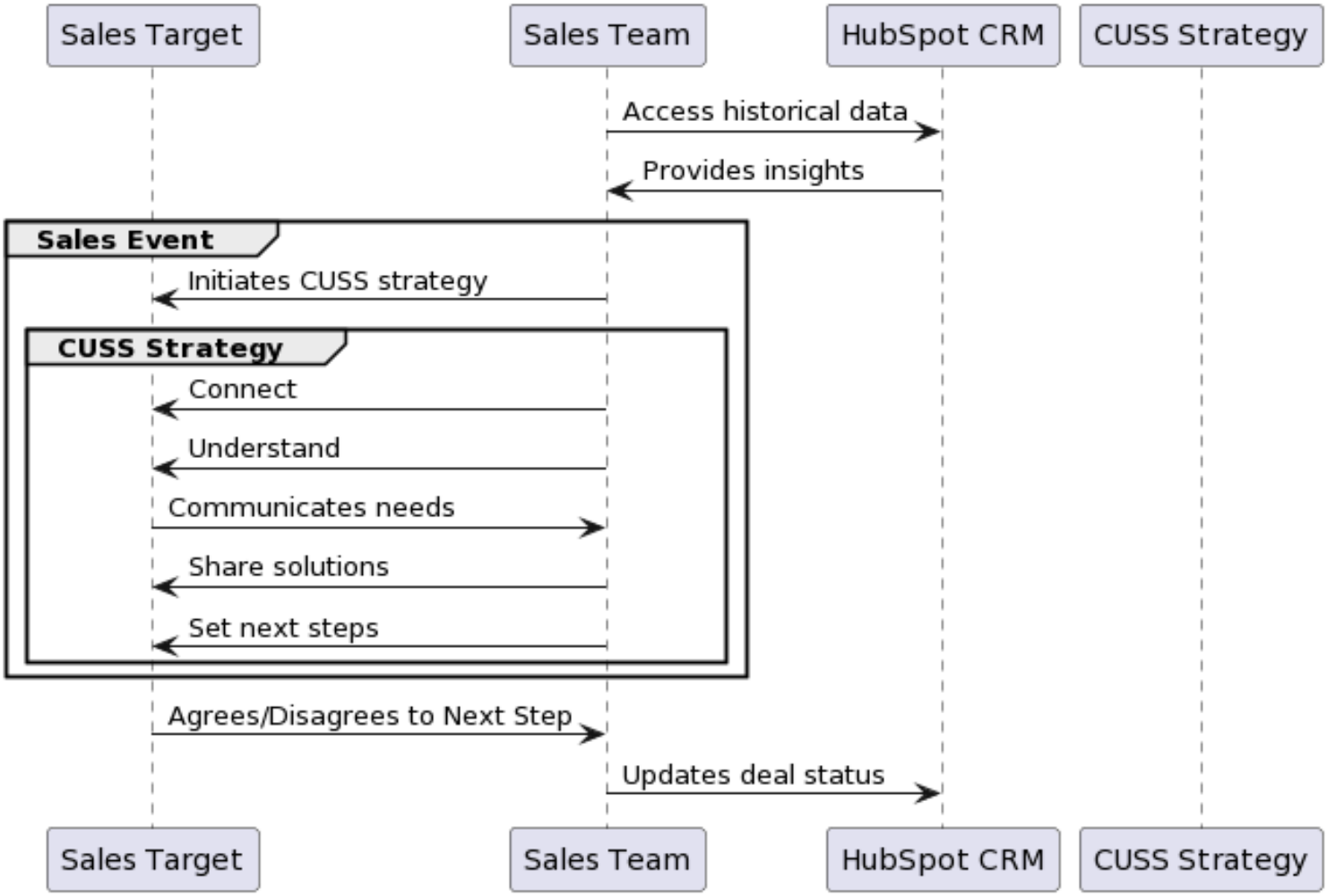
With a sufficient number of pre-trained models, a more comprehensive foundational model for StaffAny's sales team can be trained in the future using transfer learning or ensemble techniques.

Context

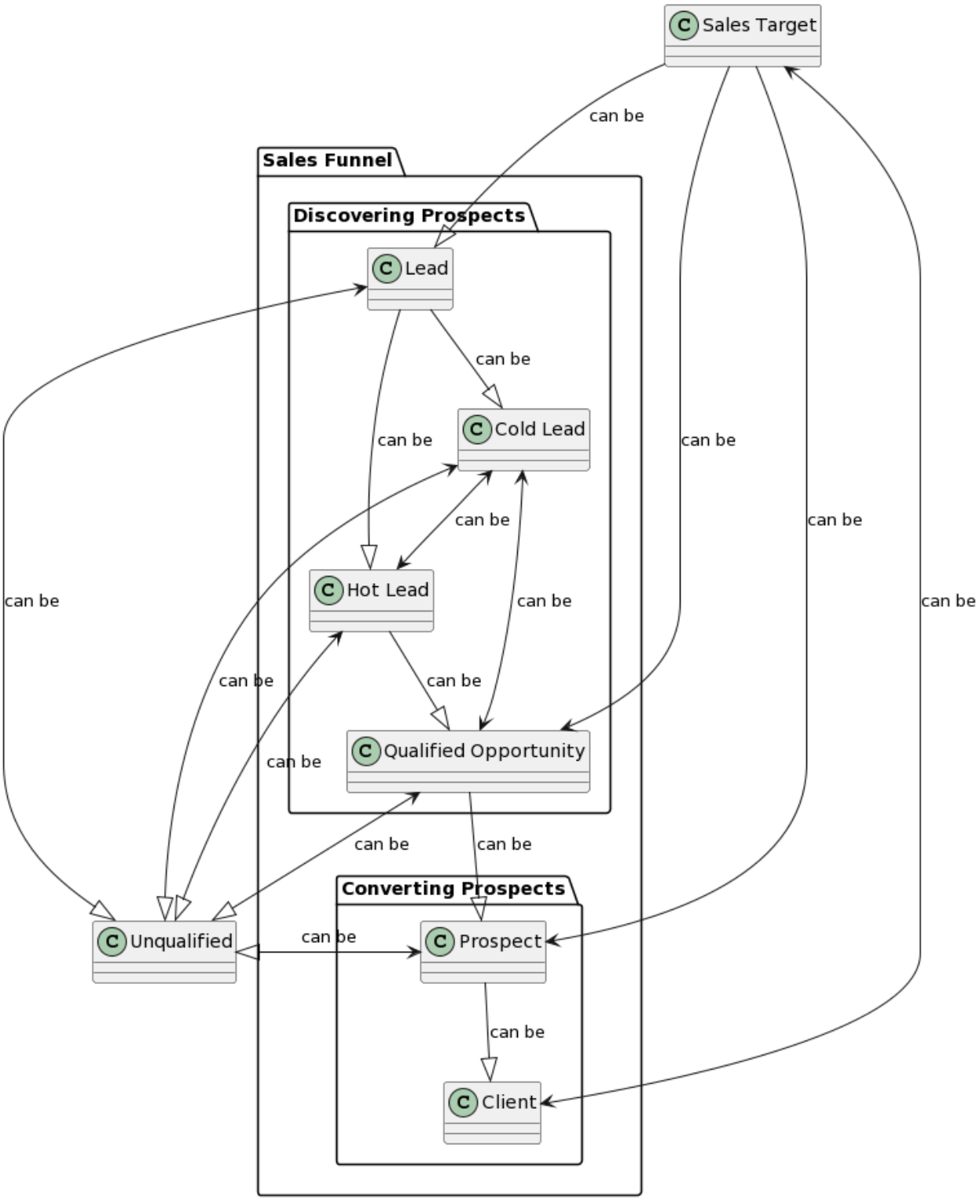
Sales operations consist of 3 key components - **sales events**, **sales funnels** and **sales cycles**.

A **sales event** involves communication between a Sales Team and their sales targets via any means. An event is defined by a 4 point communication protocol or the **CUSS strategy**,

1. **Connecting**: Building rapport with prospects.
2. **Understanding**: Learning about their needs.
3. **Sharing**: Explaining how StaffAny can meet their needs.
4. **Setting**: Planning the next steps



Each **sales event** recategorizes a **sales target**. This can be visualized as a state transition pathway from the starting *'Sales Target'* class to *'Client'* class or *'Unqualified'* class. This pathway is known as the **sales funnel**.



'Sales Target' class is the entry point and *'Unqualified'* class is the exit point of the sales funnel. *'Client'* class signifies the end of the sales funnel.

White arrowheads reflect a typical state transition pathway in the sales funnel. Represented as a series of class castings in the context of a state pattern,

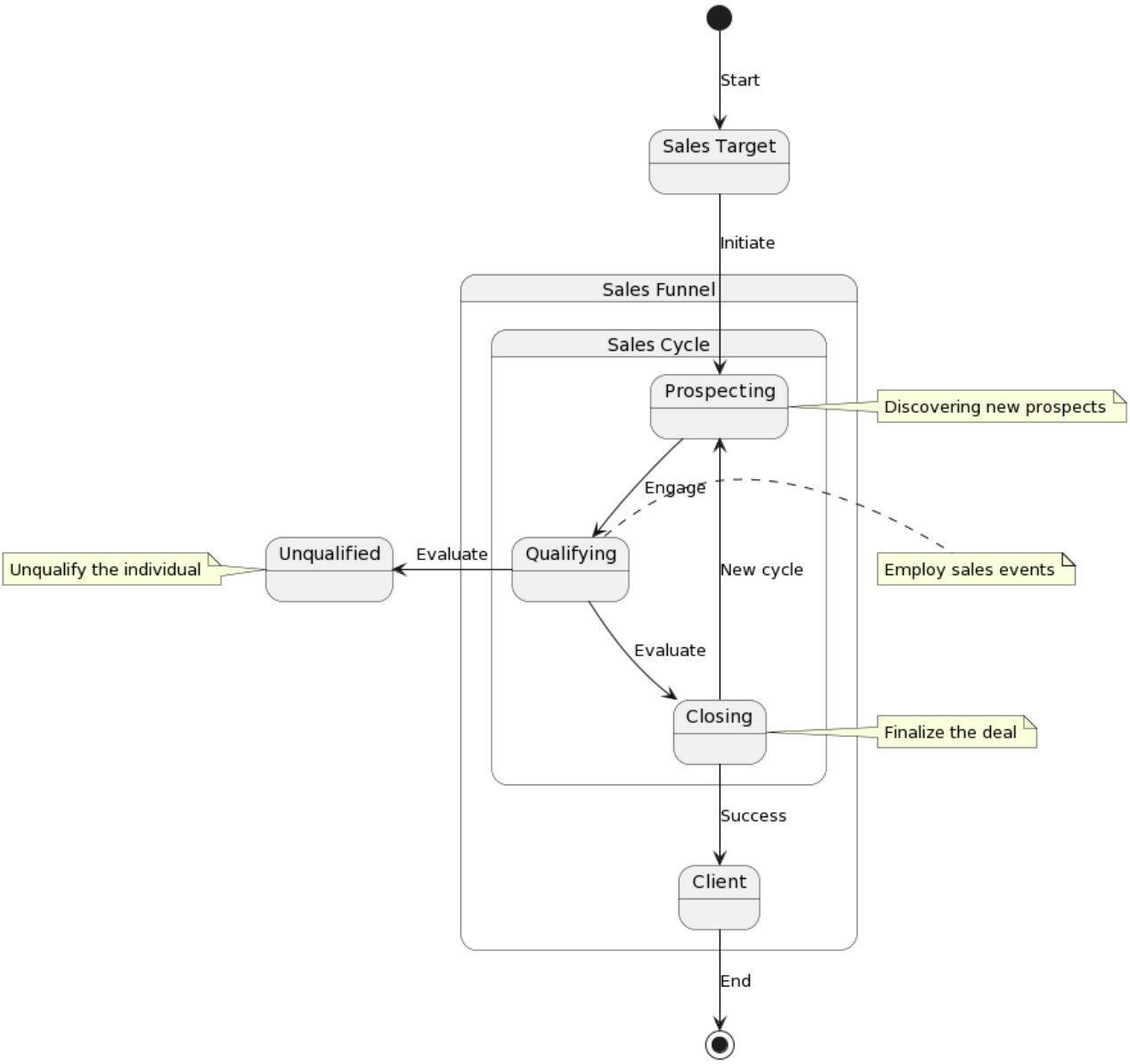
```
1. 'Sales Target' class casted to 'Lead' class.
2. If warmed up to a sales person,
      a. Then, cast to 'Hot Lead' class
      b. Else if requires more warming, cast to 'Cold Lead' class
      c. Else, cast to 'Unqualified' class
3. If strong product fit and/or likelihood for purchase,
      a. Then, cast to 'Qualified Opportunity' class
      b. Else, cast to 'Unqualified' class
4. If agree to a product demo,
      a. Then, cast to 'Prospect' class
      b. Else, cast to 'Unqualified' class
5. If purchase after demo
      a. Then, cast to the 'Client'' class
      b. Else, cast to 'Unqualified' class
```

Black arrowheads reflect atypical state transitions. Bidirectional transitions between *'Lead'*, '*Cold Lead*', '*Hot Lead*', '*Qualified Opportunity'* or *'Prospect'* classes to *'Unqualified'* class is possible. For instance, if the reasons for leaving the sales pipeline are temporary, then a transition back to any class in the sales funnel is possible once these reasons are resolved. Bidirectional transitions between *'Client'* class and '*Sales Target*' class is also possible. For instance, if service A but not service B was purchased, then 'Client' class for service A can transit to *'Sales Target'* class for service B.

'Sales Target' class can transit to any class in the sales funnel without a sales event. For instance, if an individual uses a sign up form, then it is not considered a sales event. Multiple transitions can occur in a single sales event. For instance, if an individual arranges for a product demo in a single sales event, then *'Sales Target'* class is casted to *'Lead'* class, to '*Hot Lead*' class, to *'Qualified Opportunity'* class and to *'Prospect'* class. A single transition may require multiple sales events. For instance, if an individual of the *'Qualified Opportunity'* class agrees to a product demo after 2 sales events, then 2 sales events are needed to transition to the *'Prospect'* class.

Sales events activate state transitions. Sales events are classified into 2 categories - discovering prospects and converting prospects - depending on the desired state transition.

The sales funnel can be generalized into 3 stages - Prospecting, Qualifying and Closing. This is the **sales cycle**.



The sales cycle is sequential. It must start with prospecting and end with closing. The sales cycle is cyclical. A failed or successful closing can always lead back to prospecting. The sales cycle's duration varies by prospect/client size, often longer for bigger prospects/clients.

Technology Stack

The following are the tools used for the FYP so far

Data Retrieval Toolchain			
Technology	**Type**	**Reasons**	**Status**
HubSpot CRM API	API	Customer relationship data	Will Use
Aircall API	API	Call data Call analysis	Will Use
Twilio API	API	Call data Call analysis	In Use

The Data Retrieval Toolchain employs HubSpot CRM API, Aircall API and Twilio API, and OpenAI API. They were chosen as they are the tools used by the StaffAny Sales Team. These APIs can be queried for the data needed to train the models and build any AI applications from the AI/ML pipeline.

AI/ML Pipeline Development Toolchain			
Technology	**Type**	**Reasons**	**Status**
Python	Programming Language	Rich & robust ecosystem for ML/AI applications	Using
pandas	Library	Data analysis	Using
scikit-learn	Library	ML Models	Useful

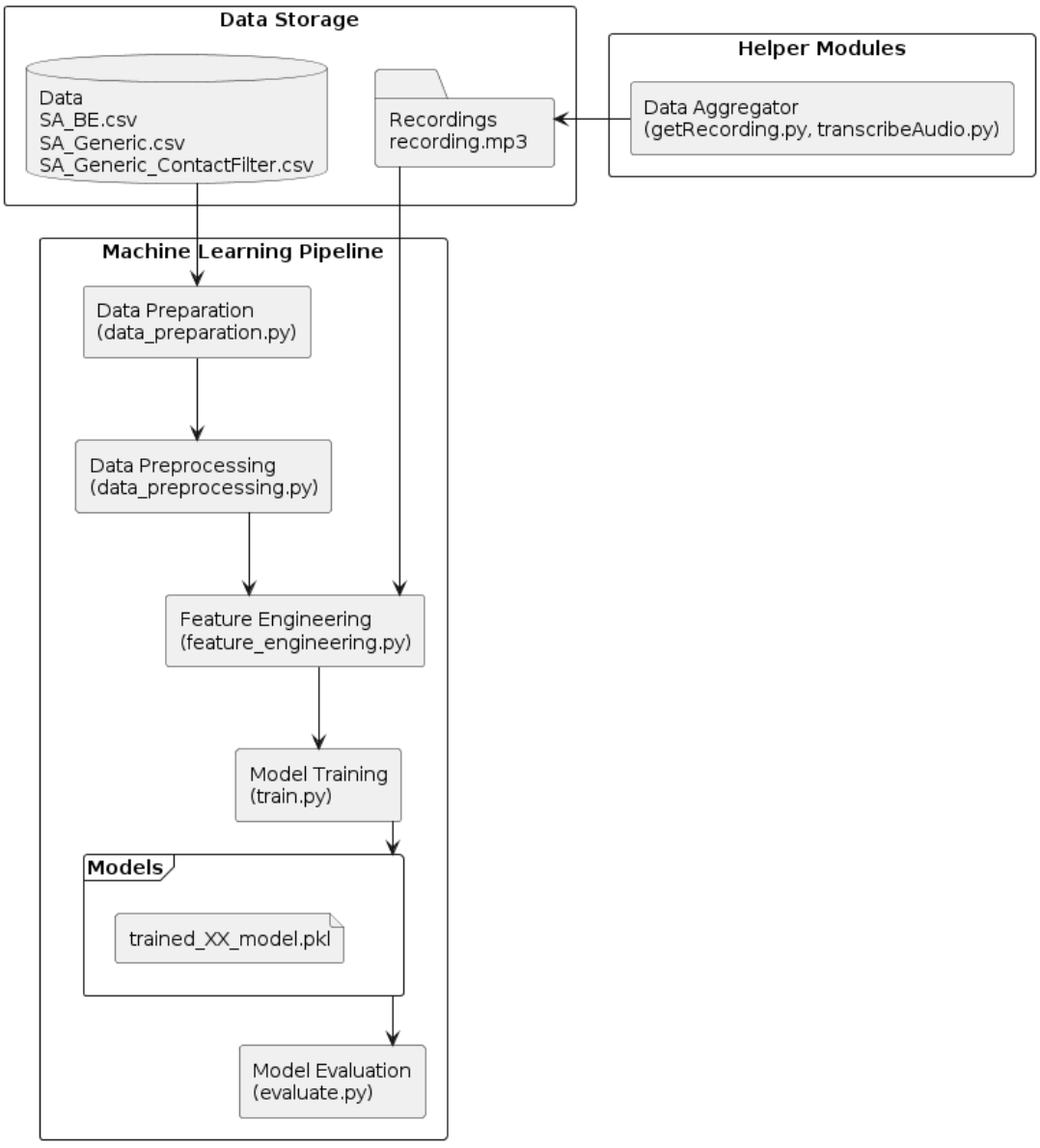
tensorflow	Library	Training and inference of deep neural networks	Useful
JupyterLab	Integrated Development Environment (IDE)	IDE focused on Data Analysis and Pipelining	Using
JupyterNotebook	Interactive Notebook	Data analysis	Using
OpenAI API	API	Data Imputation, Data Cleaning, Data Pre processing and Data Post Processing	Using

The AI/ML Pipeline Development Toolchain is centered around Python. Python was chosen for its expansive ML/AI libraries, user-friendly nature and most importantly, its alignment with StaffAny's existing tech ecosystem for data analysis. Pandas and scikit-learn are important libraries for data manipulation and straightforward ML modeling. TensorFlow supports advanced deep learning capabilities, allowing for sophisticated model training. The interactive environments of JupyterLab and JupyterNotebook are familiar tools for Exploratory Data Analysis (EDA), a key preparatory step in building the pipeline. The OpenAI API enhances the AI/ML pipeline with advanced data processing techniques via its high availability pre-trained models. This eases the burden of open-ended data analysis.

Pipeline Architecture

The ML pipeline starts with *'Data Storage'*. This holds all the essential client information and audio for processing, which serves as the raw data to be processed into the training data. In the *'Machine Learning Pipeline'*, the raw data is cleaned, standardized, and enriched with new features and labels to eventually produce the training data. The training data is then used for model training. *'Models'* are the outcome of this training, ready for making predictions or further fine-tuning. The model is then evaluated after it has been trained at the end of the pipeline

'Helper Modules' work alongside the *'Machine Learning Pipeline'*, helping in auxiliary functions like transcribing recordings that contribute to augmenting the training data with open-ended data.



Implementation

The development of the AI/ML pipeline began with requirement gathering. This helped in understanding the sales operations and its 3 key components. This was critical in problem modeling and discovering possible data captures within the current sales operations.

The next step was conducting an EDA on 2 key numeric data sets - SA_BE and SA_Generic. These 2 data sets contained data collected from customer interactions of existing clients (SA_BE) and past and present prospects (SA_Generic). This investigative phase laid the groundwork for subsequent steps. EDA helped in scrutinizing the existing data. This revealed key insights about the raw data and guided decisions on data cleaning, feature selection, and model choice. Both data sets were cleaned and preprocessed to constitute the training data set.

An EDA on sales events data was also conducted. A variety of open-ended data points were considered but finalized to 2 selections — interaction notes from HubSpot and call recordings from Twilio and Aircall. Additionally, the transcription of calls, facilitated by whisper from OpenAI API, helped to enrich the raw data. Calls devoid of substantial information were marked accordingly to maintain data integrity. The open-ended data was then transformed into structured features. This transformation distilled meaningful features based on the CUSS strategy employed during sales events that could significantly influence the predictive power of the resultant models.

The next step involved labeling data that best illustrated the state transition within the sales funnel. This was mostly done through identifying dates where the records changed their states as reflected in the state transition in the sales funnel. This was primarily achieved via the deal object obtained from the Hubspot API. Each deal object presented as a discrete object characterized by date, time, contact information, and a Hubspot Deal label that corresponded to the states in the sales funnel diagram.

Once these steps were completed, it formed the training dataset that served as input into the AI/ML pipeline, augmented by the new features and labels.

Progress

3 models were generated through the pipeline - a Linear Regression (LR) model, a Random Forest Classifier (RFC) model and a Support Vector Machine (SVM) model.

The LR model was chosen as it was great for understanding how different numeric factors in the sales operations led to a completion of a sales cycle. The RFC was chosen for predicting whether an individual will enter the next step in the sales funnel, useful for relating state transitions with the CUSS strategy. It also helped to

identify which features mattered most in the sales event, sales funnel and sales cycle. The SVM model was chosen to deal with complex and non-linear relationships in the training data. It was a suitable model choice for relating the different features within the data set.

The model was assessed using a test/training data split with a 2:8 ratio. The predicted outcomes from the test set aligned strongly with the actual results. A satisfying result was also that the prediction ability was between 70-80%. This is desirable since the danger of overfitting was also a consideration in the implementation.

The starting AI/ML pipeline has been accomplished, but no user testing on the models or additional AI applications have been built.

Timeline

AI/ML Pipeline Development Toolchain		
Week	**To Do List**	**Feature**
1-3	Research and select suitable Machine Learning models. Begin developing algorithms to rank and qualify leads. Implement Natural Language Processing (NLP) techniques to process, understand and generate business texts Integrate the qualification module with existing data sources. Begin initial testing on historical success data.	AI-based Qualification Module
4	Integrate Django with HubSpot API. Ensure seamless data flow between the AI modules and HubSpot	Integration with HubSpot
4-6	Develop & implement the Feedback Loop Module Enable the sales team to input feedback directly for user testing Test feedback system	Feedback Loop Module

	Adjust model training based on initial feedback and ensure stability of data exchange.	
7-10	Design AI-driven Retargeting Module. Use models to predict optimal retargeting times Integrate with HubSpot API and sales operation workflow to produce AI-Generated sales pitches on demand	AI-driven Retargeting Module
11-12	Begin Final Report and Presentation Prepare for Production	Finalize FYP

Future Work

Over the next 12 week period, the plan is to expand on the current AI/ML pipeline with modules specifically designed to refine lead qualification and enhance integration with StaffAny's systems. This strategic addition aims to automate and optimize the sales process, focusing on improving the above stated metrics. By employing machine learning and natural language processing techniques, higher precision in identifying and engaging with potential leads and better data flow into the AI/ML pipeline for actionable insights is anticipated.

The introduction of these modules is expected to positively impact the key sales metrics laid out in the objectives. With improved lead qualification and retargeting strategies, an increase in the total number of appointments set, a higher conversion rate from cold contacts to appointments, and more effective sales activities overall can be expected. Furthermore, by leveraging historical interaction data, the sales cycle duration can be shortened, enabling quicker transitions from initial contact to sales closure.

Conclusion

The rigorous development of an AI/ML to enhance the sales processes for StaffAny has been illustrated. From early testing results, the AI/ML pipeline shows promise in its ability to alleviate the existing problems of the sales team. However, more rigorous model selection, benchmarking and user testing is still required to completely solve the problem faced by the sales team. It is also unclear how this pipeline will eventually integrate with StaffAny's own data pipeline hosted on Google Cloud Platform. Finally, a clear 12-week plan for the integration of critical modules was also proposed. These enhancements are anticipated to improve the key metrics laid out in the objectives.

Annex

Fig.1 PlantUML

```
@startuml
participant "Sales Target" as I
participant "Sales Team" as ST
participant "HubSpot CRM" as CRM
participant "CUSS Strategy" as CUSS

ST -> CRM : Access historical data
CRM -> ST : Provides insights
group Sales Event
  ST -> I : Initiates CUSS strategy
  group CUSS Strategy
      ST -> I : Connect
      ST -> I : Understand
      I -> ST : Communicates needs
      ST -> I : Share solutions
      ST -> I : Set next steps
  end
end
I -> ST : Agrees/Disagrees to Next Step
ST -> CRM : Updates deal status
@enduml
```

Fig 2. PlantUML

```
@startuml
Class "Sales Target" {
}

Class Unqualified {
}

package "Sales Funnel" {
  package "Discovering Prospects" {
    Class Lead {
    }
    "Sales Target" --|> Lead : can be
    Lead <--|> Unqualified : can be

    Class "Hot Lead" {
    }
    Lead --|> "Hot Lead" : can be
    "Hot Lead" <--|> Unqualified : can be

    Class "Cold Lead" {
    }
    Lead --|> "Cold Lead" : can be
    "Cold Lead" <--> "Hot Lead" : can be
    "Cold Lead" <--> "Qualified Opportunity" : can be
    "Cold Lead" <--|> Unqualified : can be

    Class "Qualified Opportunity" {
    }
    "Hot Lead" --|>  "Qualified Opportunity": can be
    "Sales Target" --> "Qualified Opportunity" : can be
    "Qualified Opportunity" <--|> Unqualified : can be
  }


  package "Converting Prospects" {
    Class Prospect {
    }
    "Qualified Opportunity" --|>  Prospect: can be
    "Sales Target" --> Prospect : can be

    Class Client {
    }
    Prospect --|>  Client: can be
    Prospect <-right-|>  Unqualified: can be
    "Sales Target" <--> Client : can be
  }
}
@enduml
```

Fig 3. PlantUML

```
@startuml
state "Sales Target" as SalesTarget
state "Sales Funnel" as SalesFunnel {
  state "Sales Cycle" as SalesCycle {
    state "Prospecting" as Prospecting
    state "Qualifying" as Qualifying
    state "Closing" as Closing
  }
    state "Client" as Client
}
state "Unqualified" as Unqualified

[*] --> SalesTarget : Start
SalesTarget --> Prospecting : Initiate
Prospecting --> Qualifying : Engage
Qualifying --> Closing : Evaluate
Qualifying -left-> Unqualified : Evaluate
Closing --> Client : Success
Closing -up-> Prospecting : New cycle
Client -down-> [*] : End

note right of Prospecting : Discovering new prospects
note right of Qualifying : Employ sales events
note right of Closing : Finalize the deal
note left of Unqualified : Unqualify the individual
@enduml
```

Fig 3. PlantUML

```
@startuml
skinparam linetype ortho

rectangle "Data Storage" {
     database  "Data\nSA_BE.csv\nSA_Generic.csv\nSA_Generic_ContactFilter.csv"
as Data
    folder "Recordings\nrecording.mp3" as Recordings
}

rectangle "Machine Learning Pipeline" {
    agent "Data Preparation\n(data_preparation.py)" as DataPrep
    agent "Data Preprocessing\n(data_preprocessing.py)" as DataProc
    agent "Feature Engineering\n(feature_engineering.py)" as FeatEng
    agent "Model Training\n(train.py)" as ModelTrain
    agent "Model Evaluation\n(evaluate.py)" as ModelEval
    frame "Models" {
        file "trained_XX_model.pkl" as LR
    }
}

rectangle "Helper Modules" {
    agent "Data Aggregator\n(getRecording.py, transcribeAudio.py)" as DataAgg
}


Data --> DataPrep
DataPrep --> DataProc
DataProc --> FeatEng
FeatEng --> ModelTrain
ModelTrain --> Models
Models --> ModelEval

DataAgg -left-> Recordings
Recordings --> FeatEng

@enduml
```

Company Work Breakdown

Objectives:	
StaffAny Aims to help (Who - Persona)	Revenue Team
Achieve (What - Desired outcome)	A system that automates the hot lead generation process, allowing Account Managers to focus more on lead qualification & improving the sales lifecycle.
Because (Why - current problem)	The sales unit struggles with a time-consuming sales lifecycle despite using tools like HubSpot and ChatGPT UI. This diverts focus from crucial tasks such as refining sales pitches, closes & lead qualification. To address this, we plan to employ AI and automation, to improve & shorten this sales lifecycle and as a stretch goal leverage Database APIs and search engine SDKs, streamlining research & cold lead generation.
Why Now	This comes at a time where StaffAny faces less Account Executive's that are able to drive revenue growth for the company. As a result, the time spent of a single Account Executive is crucial to the company and ideally, we want the time spent to be spent predominantly on driving revenue growth and not on auxiliary sales functions like lead generation
We aim to solve it via (How - Preliminary Idea)	Improving Appointment Set Number per same call activity Improving Conversion rate per same appointment met activity Improving Sales Life Cycle duration (reduce touch point per contact)

Key Results:	
Company should understand where we are as a product	Data cleaning, Data Labeling stage (Refining a primary dataset as test dataset)
Improvement Metric	Increasing Appt Set Number by 2 every month

Timeline

Phase 1:	20th Feb - Cleaned data to feed the ML Model 28th Feb - ML Pipeline ARchitecture to be drawn up

RACI:	Overall	Engineering Lead	Data Compliance	Data Preparation	Pipeline Management
[R]esponsible	Eugene	Jeremy	Karl?	Damba/Anis	Damba/Anis
[A]ccountable aka Directly Responsible Individual	Michael	Michael	Michael	Michael	Michael
[C]onsulted	Eugene, Janson (formal Approval)	Jiayi Kishan	-	-	Jeremy
[I]nformed	Jeremy, Kaiyi	Eugene	Janson	Eugene	Eugene