AI Product Manager
Capstone Project Proposal

Conversational AI in IVR Customer Service Banking Application

MICHAEL KWONG HIN SANG

Business Goals

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you're labeling images, how will this help the business?	Automated Interactive Voice Response (IVR) systems are used in many banks today to serve millions of customers with inquiries such as account balances and transactions. IVR systems were introduced decades ago and often provide poor customer experience (CX). Customers navigate several hierarchical menu options and select static options by keying in touch tone DTMF keys to go from one dialog module to the next. This frustrates customers and ML/AI can provide value for companies to establish better customer service experience and with customers being their most valuable asset! Conversational AI can understand customers speaking their requests in a free form interaction rather than selecting static choices from different menus. AI/Natural Language Processing (NLP) will analyze the voice conversations and understand the context and content of what the caller says, and take caller directly to the IVR function to self-serve their requests.

Business Case Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success.	By introducing Conversational AI, the customers interactions with the bank will be better. The customer will not have a rigid and robotic conversation with the traditional IVR, but instead Conversational AI will understand the context and content of the discussion at all times. For example, if a caller wants to make a change in payment amount, s/he does not need to start all over but just change the amount as if when talking with a live person. By allowing users to speak normally and change their minds, Conversational AI provides a more friction-free interface between customers and their banks. This will increase customer satisfaction, increase company brand, increase self-service transactions in the IVR application, and reduce call transfer escalations to agents to only the more complex interactions. The average handle time (AHT) of calls served by agents can be reduced and save the company in labor cost reduction. For a huge bank with millions of calls per day, this can amount to millions of dollars of savings per year.
Application of ML/AI What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?	The task I will use AI is Machine Learning/Natural Language Processing (NLP). The Natural Language Understanding (NLU) language model created as part of NLP will replace the hierarchical static menus options in the current IVR. AI will analyze what the caller is saying: NLU will be done by identification of named entities and word patterns through tokenization, stemming, syntax and lemmatization. By analyzing the words, AI can find the semantic tag for what the caller inquiry is and capture the semantics of the caller utterance. AI will return a statistical language modelling (SLM) with the predicted intent and continue with the next action in the IVR. This will help self-service the customer faster and increase customer satisfaction when using the IVR application without transferring to a customer service agent.

	The NLU model produced by AI/NLP will understand the context, intent and the sentiment of the text. Sentiment analysis will help to understand customer speech and identify issues and help customers more quickly before issues even escalate. This will ultimately increase customer satisfaction.

Success Metrics

Success Metrics What business metrics will you apply to determine the success of your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison.	To determine the success of my product, I will look at the following success metrics. For each, I will have a baseline value of: -The current IVR percentage of call transfers to agents per month. -The current IVR percentage of call hang-ups per month. -The current IVR percentage of functions where callers self-served after choosing several menu options. After the new product with Conversational AI is implemented in production, I will collect the following future success metrics: -The future IVR percentage of call transfers to agents per month with AI. -The future current IVR percentage of call hang-ups per month with AI. -The future IVR percentage of the functions self-served in IVR with caller's intents predicted correctly with AI/NLP/NLU. In this way, I will compare the percentages with the baseline and determine how successful the new product will be.

Data

Data Acquisition Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?	The data set will be sourced from a corpus of raw audio data of customers' spoken requests in the banking industry. These utterances will be captured by an Automatic Speech Recognizer (ASR) and translated from speech raw audio to texts. Recurrent Neural Network (RNN) will be used in the speech recognizer to recognize and predict the sequences of text. The texts will then be annotated with semantic tags of the intents corresponding to what callers will request with Conversational AI. The labelled corpus of training dataset will contain an estimation of at least 100,000 sentences to classify the intents. It will require supervision by human annotators, and training of the NLU model. The dataset will be balanced for all possible intents in the IVR menus so that the model does not skew to any predicted intent. We will start with a corpus of data from a vendor so that the cost and time to obtain a cleaned corpus of data is less. In the banking industry there are strict regulations about personally identifying information (PII) such as account numbers, social security, PIN. Data governance is key and proper disclosures will inform callers. The PII information from customer will not be saved when the Conversational AI will be integrated to the IVR. The caller will already be authenticated prior to reaching the Conversational AI. After implementation, the data will be ongoing to continuously tune the AI model. The Conversational AI will be capable of improving its proficiency based on the new data inputs it receives from more caller's spoken utterances. When it cannot comprehend what a caller is saying, an agent will take over seamlessly. The AI system will then store the caller's utterances for analysis so that when a similar question or query is posed in the future, it can handle it without human intervention.
Data Source	

Consider the size and source of your data; what biases are built into the data and how might the data be improved?	The data source will account for biases. The model itself can only learn based on the input data provided by humans. The training data will not ignore some target group, accents, minorities. The date will have diversity in the training set. There will be enough unbiased data as well as diverse annotators who will have an eye if there is unconscious bias caused by humans during annotation. This is important for strictly regulated banks to not have biased in AI model that would otherwise lead to loss of reputation and even lawsuits. The training audio data source will be collected from mobile devices, as well as land lines to have more real-life situations. Training data that are only from mobile audio with a great deal of noise will not create a good unbiased model. So, we will collect more balanced training data to get a better model. The data will be improved by collecting more training data on an ongoing basis for the intents classes that are lacking. This will prevent bias towards any class.
Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels versus any other option?	The dataset will be transcribed from speech and then annotated with the labelled intents corresponding to the functions that are available from the IVR menus. For example, after authentication, callers hear a menu and key in dtmf-1; followed by another menu and key in dtmf-2; followed by yet another menu and key in dtmf-5 up to the password reset function they want. However, when using AI, callers will simply say ‘password reset’ after authentication and go directly to password reset module to self-serve their requests. Some of the utterances that cannot be translated to the IVR functions will be categorized to an agent intent. After sentiment analysis of calls, some calls will also be tagged for immediate escalation. These agent and escalation intents will transfer out to special agent groups skilled to serve these customers the best. I chose this labelling scheme because these are the options of the functions available in the IVR menus.

	Instead of callers keying in several dtmf keys and listening to whole menus, they can directly go to the functions they want after the labels/intents are returned from the conversational AI/NLU.

Model

Model Building How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why?	To build the model, we will partner with Nuance vendor. We will use a generic well-trained bank model and adopt transfer learning instead of training the model from scratch. We will optimize our NLU model and convert unstructured sentences utterances as inputs and convert them into structured data format as output intents labels. The trained statistical language model (SLM) will be outsourced to the vendor and integrated in the IVR call-flow after callers are authenticated. The model itself will not contain sensitive account information but return possible semantic tags translated to intents class labels which the IVR Dialog modules will then process to assist the customer.
Evaluating Results Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?	For NLU model predictions, I will measure the success with a confusion matrix. It will show the predicted intents versus the expected intents. For example, when callers say “I want password reset”, the NLU should predict “password” intent but was that the prediction? To measure that, the results will be in a confusion matrix with the numbers in diagonal increasing when the expected intents (rows) match the predicted intents (columns). Below is a just a sample example for illustration.

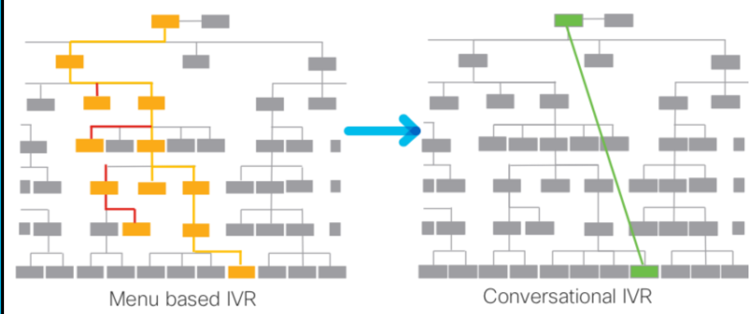
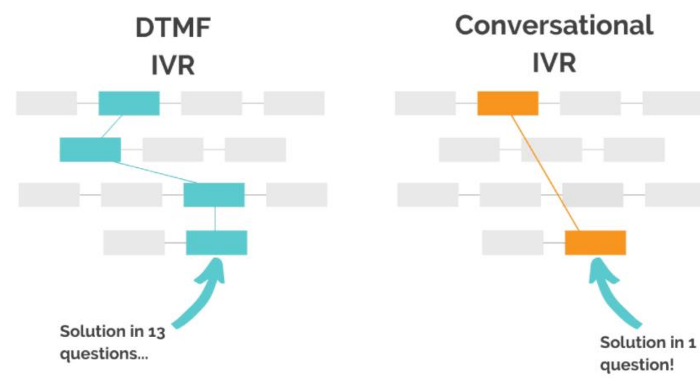
Confusion matrix

	account_balance	contact	disposable	interest_rate	loans	mobile_banking	mortgage	overdraft	pay_person	payday	savings_target	smalltalk.name	smalltalk_hello	smalltalk_thank	spend_category	statement	upcoming_bills
account_balance	11	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
contact	0	9	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
disposable	1	0	13	0	1	0	0	0	0	0	0	0	0	0	0	0	0
interest_rate	0	0	0	9	0	0	0	1	0	0	0	0	0	0	0	0	0
loans	0	0	0	0	12	0	1	0	0	0	0	0	0	0	0	0	0
mobile_banking	0	0	1	0	0	13	0	0	0	0	0	0	0	0	0	0	0
mortgage	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
overdraft	0	0	0	1	0	0	0	13	1	0	0	0	0	0	0	0	0
pay_person	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0
payday	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
savings_target	0	1	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0
smalltalk.name	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0
smalltalk_hello	1	0	0	0	0	0	0	0	0	1	0	0	10	0	0	0	0
smalltalk_thank	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0
spend_category	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9	0	0
statement	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	10	0
upcoming_bills	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	10

I will refine the training data and train the model and repeat until satisfied with the F1-score. When training, I will look at increasing F1-score to be very close to 1.0. This will then mean the data are resolving perfectly by the NLU for the intents callers will ask for, and the performance of NLU is high.

Minimum Viable Product (MVP)

Design What does your minimum viable product look like? Include sketches of your product.	In the Minimum Viable Product (MVP), the callers no longer need to navigate long hierarchical menus with DTMF/touch tones keys to go to the self-serve function they called for. The MVP will have Conversational AI so that callers can directly speak the reason they are calling for. They can then go directly to the self-service function without navigating all the menus in IVR as shown in the sketches below.

	 
Use Cases What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product?	The design is for the persona of all the bank customers calling the tollfree number for customer service. This includes diverse customers of all background and age appropriate that have a business relationship with the bank: credit card, student loan, mortgage, car insurance, brokerage account, loans, etc. The major epic-level use cases are for self-serving IVR functions efficiently and accurately without navigating several menus and options. Customers will access the product after calling the bank's customer service main phone number and authenticate with their bank account information. Customers will speak the reason they are calling. The AI model will return the intents and next go to the functionalities they called for.
Roll-out	The roll-out will have a pre-launch and post-launch plan.

How will this be adopted? What does the go-to-market plan look like?	In the pre-launch plan, we will obtain signoff from all stakeholders, including Quality Assurance and Model Validation teams confirming that the different product functions use cases are passed successfully. In the post-launch plan, the customers will be notified after the initial welcome message to the bank that, to serve them better, we have upgraded our customer service application. They will be guided to speak what they are calling for, instead of navigating a series of menu options.

Post-MVP-Deployment

Designing for Longevity How might you improve your product in the long-term? How might real-world data be different from the training data? How will your product learn from new data? How might you employ A/B testing to improve your product?	Post-MVP-Deployment, the Conversational IVR will keep learning from the conversations. For example, when a caller says something s/he does not understand, Conversational AI will guide the caller to an agent. But the next time when it is a similar query, the AI will be capable to handle it and self-serve in the IVR, as these inquiries and utterances will be fed to the model and continuously and actively trained from the new data. Each roll-out will have the model versioned in Github tool. After roll-out, the v1 controlled model will create real-world production data from callers calling in different environment such as with background noises working from home, with kids noise, or TV playing in the background. The calls can also be from cell phones or Skype Voip calls, from car speakers, etc. The ASR will capture these customer calls with speech-to-text (STT). The texts will be transcribed, annotated, and then used to re-train the new model to be more real-life. Production traffic will be split with 20% of calls sent to the challenger v2 model and 80% to a well-tested v1 controlled model. The percentage will be controlled from

	the cloud by the F5 load balancing network that monitors incoming distribution of phone calls. A/B testing can help to get some good experimental production data and see if the newer version v2 challenger model would work better. And if so, we will switch all the 100% user traffic to the new v2 challenger model. After each release, we will actively continue to measure the performance of the model and update as necessary.
Monitor Bias How do you plan to monitor or mitigate unwanted bias in your model?	I plan to monitor the following unwanted biases in my model: 1. Algorithm Bias Caused by how the model was developed or how the model was trained that results in unfair outcomes. How to monitor or mitigate: I will use A/B testing on a few models to check for biases. 2. Exclusion Bias Data is scrubbed and pre-processed before being used in training or testing a machine learning model. This cause exclusion bias. It occurs when we remove or add features that we think are not relevant or relevant such as skewed data on only some accents but not diverse. How to monitor or mitigate: I will bring in domain knowledge experts to conduct feature engineering i.e. to find out which should be included and not included in the dataset. 3. Measurement Bias Differences in the data collected for training differs from the data collected during production. How to monitor or mitigate: I will do Quality control on the data being collected. I will also plan to use production data after roll-out to continuously train model and improve it.