

Capstone Project Proposal



Osama Alsubaie

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?

Industry problem:

Individual fraud is a growing problem in the banking sector in Saudi Arabia. Fraudsters often pose as bank employees or representatives of financial institutions in order to trick victims into revealing their personal and financial information. They may also use social engineering techniques to gain the victim's trust, such as pretending to be a customer service representative who needs to verify the victim's account information.

All banks in Saudi Arabia are using 2-step verification after signing into the bank system with the account information or trying to buy something online after entering the credit card info. Fraudsters are increasingly using two-factor authentication (2FA) to trick victims. They will call the victim and pretend to be from the bank or a financial institution. They will then ask the victim for their 2FA code, which is a temporary code that is sent to the victim's phone in order to verify their identity. Once the fraudster has the 2FA code, they can use it to access the victim's account and withdraw money or make unauthorized purchases.

Despite the fact the SAUDI CENTRAL BANK (SAMA) has warned the public about the dangers of individual fraud and has issued a number of tips to help people protect themselves, many people are getting frustrated again and again so how we can benefit from the AI in this issue.

Why use ML/AI:

Machine learning (ML) and artificial intelligence (AI) can be used to solve the problem of individual fraud in the banking sector in a number of ways.

Our project will focus on **Identifying the fraud transaction based on the victims' behavior and their common qualities.** ML algorithms can be trained on data from past fraud cases to identify fraud transactions based on the victims' behavior and qualities. This information can then be used to develop **new fraud transaction detection models that can be used to identify potential fraud transactions and block them.**

	<div>How ML/AI can provide value:</div> <p>In our case, we want to identify the fraud transactions based on the victims' behavior and qualities. There are many ML algorithms that can help us in this case, decision tree or random forest algorithm. The use of these algorithms will help us to identify the possible fraud transactions based on the input data given to the model and block them.</p>						
<div>Business Case</div> <p>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.</p>	<p>ML and AI can provide significant value to the banking sector by helping to prevent individual fraud. By identifying patterns of the victims' behavior, ML and AI can help banks to protect their customers from fraud and financial loss.</p> <div>ML and AI can also help banks to:</div> <div> <div>1- Improve customer satisfaction:</div> <p>By preventing fraud, banks can help to improve customer satisfaction. This is because customers are more likely to be satisfied with their bank if they feel confident that their personal and financial information is safe.</p> <div>2- Reduce costs:</div> <p>By preventing fraud, banks can reduce their costs. This is because fraud investigations and losses can be costly for banks.</p> <div>3- Increase market share:</div> <p>By being seen as a leader in fraud prevention, banks can attract new customers and increase their market share.</p> </div>						
<div>Application of ML/AI</div> <p>What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?</p>	<table> <tr> <td>Task</td><td>Identify patterns of the victims' behavior and their common qualities such as [age, gender, location, educational background, etc....]</td></tr> <tr> <td>Description</td><td>Machine learning (ML) algorithms can be trained on data from the user's transaction history, behavior, and qualities to compare it with the current transaction. This data can include things like the user's spending habits, the devices they use to make transactions, and their location. The ML algorithm can then identify patterns and anomalies that may indicate fraud.</td></tr> <tr> <td>Achievement</td><td>This information can then be used to develop new fraud transactions detection models that can be used to identify fraud transactions and block them. By using ML algorithms, the bank can improve their fraud detection capabilities and protect themselves from financial losses, also make the customer experience better.</td></tr> </table>	Task	Identify patterns of the victims' behavior and their common qualities such as [age, gender, location, educational background, etc....]	Description	Machine learning (ML) algorithms can be trained on data from the user's transaction history, behavior, and qualities to compare it with the current transaction. This data can include things like the user's spending habits, the devices they use to make transactions, and their location. The ML algorithm can then identify patterns and anomalies that may indicate fraud.	Achievement	This information can then be used to develop new fraud transactions detection models that can be used to identify fraud transactions and block them. By using ML algorithms, the bank can improve their fraud detection capabilities and protect themselves from financial losses, also make the customer experience better.
Task	Identify patterns of the victims' behavior and their common qualities such as [age, gender, location, educational background, etc....]						
Description	Machine learning (ML) algorithms can be trained on data from the user's transaction history, behavior, and qualities to compare it with the current transaction. This data can include things like the user's spending habits, the devices they use to make transactions, and their location. The ML algorithm can then identify patterns and anomalies that may indicate fraud.						
Achievement	This information can then be used to develop new fraud transactions detection models that can be used to identify fraud transactions and block them. By using ML algorithms, the bank can improve their fraud detection capabilities and protect themselves from financial losses, also make the customer experience better.						

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.

The success metric is the bank customer experience and the cost deduction. These two metrics clearly defined and can easily be measured.

The baseline is important because it provides a starting point for measuring the performance of the fraud transaction detection system. By comparing the performance of the system to the baseline, you can see how much improvement the system has made.

The product is trained on historical data of fraudulent transactions. The baseline for this system is the number of fraudulent transactions that were detected by the manual and rule-based fraud detection systems before the machine learning model of our product was implemented.

The product also improves the customer experience by reducing the number of false positives. False positives are transactions that are incorrectly identified as fraudulent. This can be a frustrating experience for customers, as they may have to spend time and effort to prove that they are not committing fraud. The product reduces the number of false positives by using a machine learning model that is able to distinguish between fraudulent and legitimate transactions with greater accuracy than humans.

By improving the fraud detection performance and customer experience, the product can help organizations to protect themselves from fraud and to provide a better experience for their customers.

Data

<h2>Data Acquisition</h2> <p>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?</p>	<h3>Where will you source your data from?</h3> <p>The data source for the AI system will be the bank's own dataset of fraud cases. This dataset will contain information such as the victim's name, account number, transaction history, and contact information.</p>	
	<h3>What is the cost to acquire these data?</h3> <p>The cost of acquiring this data is negligible, as the bank already has it in their data warehouse.</p>	
	<h3>Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome?</h3> <p>There are a number of personally identifying information (PII) and data sensitivity issues that will need to be addressed when using this data. For example, the data will contain the victim's name, account number, and contact information. This information is sensitive and should be protected from unauthorized access. Additionally, the data may contain information about the fraudster, such as their IP address or email address. This information should also be protected from unauthorized access, as it could be used to commit further fraud.</p>	
	<h3>Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?</h3> <p>The data will become available on an ongoing basis, as new fraud cases are reported to the bank. The data will need to be refreshed on a regular basis to ensure that the AI system is up-to-date on the latest fraud trends.</p> <p>The data will need to be anonymized before it is used to train the AI system. This means that the victim's name, account number, and contact information will be removed from the data. The data will also be encrypted to protect it from unauthorized access.</p>	
	<p>By addressing these PII and data sensitivity issues, we can ensure that the AI system is used ethically and responsibly.</p>	
	<h2>Data Source</h2> <p>Consider the size and source of your data; what biases are built into the data and how might the data be improved?</p>	<p>As mentioned above, the data source for the AI system will be the bank's own dataset of fraud cases. This dataset will contain information such as the victim's name, account number, transaction history, and contact information.</p> <p>The size of the dataset depends on the number of fraud cases. however, based on the bank clients which are 18,000,000 the size will be approximately as 20,000 records.</p>

	<p>There are a number of biases that may be built into the data, For example:</p> <ul style="list-style-type: none"> - The data may be biased towards recent fraud cases. This is because banks are more likely to report recent fraud cases to the authorities. - The data may be biased towards large fraud cases. This is because large fraud cases are more likely to be investigated by the authorities. - The data may be biased towards certain demographics, such as age, gender, or race. The data may also be biased towards certain types of fraud, such as credit card fraud or wire fraud. <p>By taking these factors into account, we can better understand the biases in the data and make more informed decisions about how to improve the data.</p> <p>The data can be improved by:</p> <p>1- Diversifying the data: The data can be diversified by including more records from different demographics and different types of fraud. This will help to reduce the bias in the data.</p> <p>2- Cleaning the data: The data can be cleaned by removing any outliers or errors. This will help to improve the accuracy of the AI system.</p> <p>3- Enriching the data: The data can be enriched by adding additional information, such as the victim's location or the fraudster's IP address. This will help the AI system to make more informed decisions.</p>
<p>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?</p>	<p>The labels that I decided to add to the data are:</p> <ol style="list-style-type: none"> 1- Fraudulent: This label indicates that the transaction was fraudulent. 2- Not Fraudulent: This label indicates that the transaction was not fraudulent. 3- Other: This label indicates that the transaction is not yet known to be fraudulent or not fraudulent. <p>I decided on these labels because they are the most common labels used for fraud detection. These labels are also easy to understand and interpret.</p> <p>Other options for labels, such as:</p> <ol style="list-style-type: none"> 1- High Risk: This label indicates that the transaction is at high risk of being fraudulent. 2- Low Risk: This label indicates that the transaction is at low risk of being fraudulent. <p>However, I decided against these labels because they are not as clear-cut as the labels that I chose. For example, what does "high risk" mean? Is it 50% likely to be fraudulent? 75% likely to be fraudulent? It is also difficult to determine the risk of a transaction without knowing more about the transaction.</p>

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?

The bank has already a strong and secure infrastructure and a team of data scientist and ML engineering in the digital transformation sector. The team already has an experience with building and deploying machine learning models, so I will build the model using an in-house team.

By building the model in-house, we can ensure that the model is accurate, scalable, and meets the specific needs of the bank. We can also make changes to the model as needed. This will help to ensure that the model is effective in detecting fraud.

Here are some factors we considered when deciding whether to build the model in-house or outsource it:

- 1- Cost: Building the model in-house can be more expensive than outsourcing it. However, we will have more control over the model development process.
- 2- Time to market: Building the model in-house can take longer than outsourcing it. However, we will have more flexibility in the model development process, and we can reuse this model in other bank businesses.
- 3- Security: WE need to make sure that the data is secure in order to build the model in-house, luckily that the bank has strong and secure infrastructure so we can get benefit of this.

Evaluating Results

Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?

The following model performance metrics are appropriate to measure the success of my model:

True Positive Rate (TPR):

The TPR is the percentage of actual fraud cases that are correctly identified by the model. A high TPR indicates that the model is good at identifying fraud cases.

False Positive Rate (FPR):

The FPR is the percentage of non-fraud cases that are incorrectly identified by the model as fraud cases. A low FPR indicates that the model is good at avoiding false positives.

Precision:

Precision is the percentage of victims that are identified by the model as fraud victims who are actually fraud victims. A high precision indicates that the model is good at avoiding false positives.

Recall:

Recall is the percentage of fraud victims that are actually identified by the model as fraud victims. A high recall indicates that the model is good at identifying fraud victims.

The level of performance required will depend on the specific needs of the bank. **However, a good target for the TPR and FPR is 90% or higher. A good target for the precision and recall is 80% or higher.**

The bank shouldn't rely on the model and should continue raise client's awareness by sending awareness messages as an example.

Minimum Viable Product (MVP)

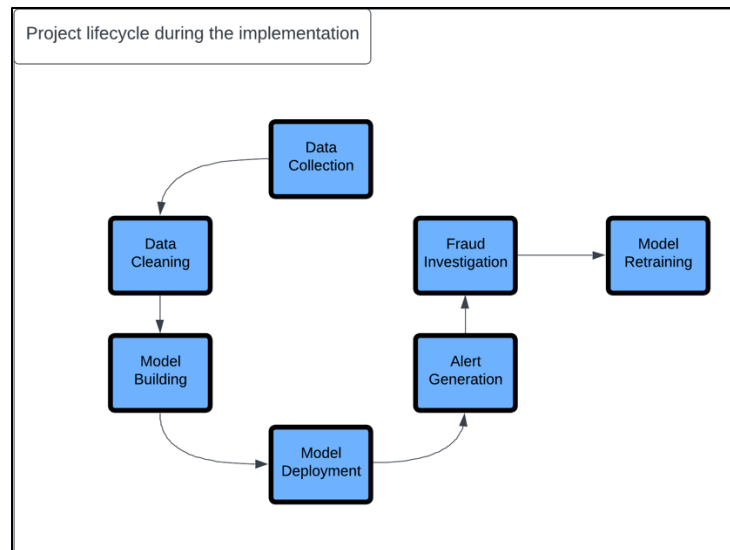
Design

What does your minimum viable product look like? Include sketches of your product.

The product is a model that will be integrated with the bank authentication system to help in classifying fraud cases and blocking them.

The bank has thousands of transactions every minute, if the system detects fraud cases, then the transaction will be blocked and a call by an agent/autoresponder to the clients will be done to check if the transaction is fraudulent or not.

After that, the data of the transaction shall be updated based on the user feedback, and the model fed with the new data to be retrained.



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?

What persona are you designing for?

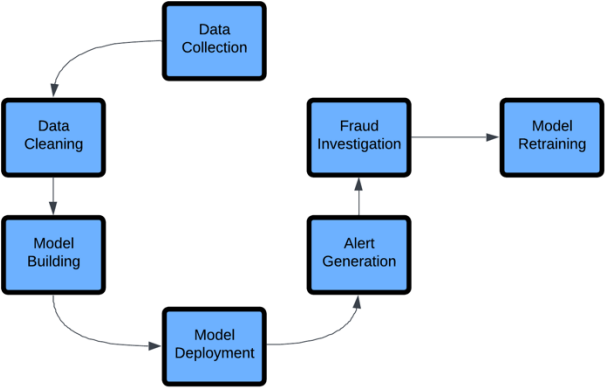
The product is designed for a bank fraud analyst who is responsible for detecting and preventing fraud. They are under pressure to detect fraud quickly and accurately, and they need a tool that can help them to do this.

Can you describe the major epic-level use cases your product addresses?

Fraud detection: The product will help to detect fraudulent transactions. This will help the bank to protect its customers from financial loss and to maintain its reputation.

User convenience: The product will be easy to use. This will make it more likely that the fraud analyst will actually use it.

Scalability: The product will be scalable. This will allow me to add new features and to support a growing the coverage of the new fraud techniques.

	<div data-bbox="703 212 1105 239" data-label="Section-Header">How will users access this product?</div> <div data-bbox="703 247 1373 346" data-label="Text"><p>The product will be integrated with the bank authentication system, the system will notify the analyst in case of any fraud detection and will block the transaction from being completed.</p></div>
<div data-bbox="204 415 328 447" data-label="Section-Header">Roll-out</div> <div data-bbox="204 487 656 592" data-label="Text"><p>How will this be adopted? What does the go-to-market plan look like?</p></div>	<div data-bbox="691 415 1404 564" data-label="Text"><p>The go-to-market plan for the product as follow: After building the system the model will be export and integrate with the bank authentication system to start analyzing the transactions. The next image shows the implementation lifecycle of the product:</p></div> <div data-bbox="695 598 1417 1144" data-label="Diagram"><div data-bbox="708 613 1052 655" data-label="Caption"><p>Project lifecycle during the implementation</p></div><pre>graph TD; DC[Data Collection] --> DataCleaning[Data Cleaning]; DataCleaning --> MB[Model Building]; MB --> MD[Model Deployment]; MD --> AG[Alert Generation]; AG --> FI[Fraud Investigation]; FI --> MR[Model Retraining];</pre></div>

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?	Continuously update the model: The model should be continuously updated with new data to improve its accuracy. This will help to ensure that the model is always up-to-date with the latest fraud trends. Add new features: Adding new features to the product will make it more powerful and versatile. For example, add new feature that allow the product to detect different types of fraud, such as credit card fraud and identity theft. Since the training data will be from the bank own dataset, then there will be no difference between the Real-world data and the training data. As a result, the model will be accurate with the real-world data. The A/B testing can be done until we achieve high Performance metrics, testing against statistically significant sample size and running tests long enough to capture any seasonality effects.
Monitor Bias How do you plan to monitor or mitigate unwanted bias in your model?	Once the model is built, it is deployed. The model is then used to analyze new data to identify potential fraud cases. If the model identifies a potential fraud case, it will generate an alert. The alert will be sent to a human analyst who will investigate the case further. The fraud detection system is an ongoing process. The model is continuously being updated with new data to improve its accuracy. The human analysts are also continuously learning and improving their skills at investigating fraud cases. The fraud detection process is also not always perfect. So the next two strategy will minimize the unwanted bias in our product. Feedback Loops: We maintain a feedback loop with our users and analysts. They play an essential role in identifying and reporting instances of bias that they encounter. This iterative feedback process enables us to continuously improve the model's performance and fairness. Ongoing Model Updates: Our fraud detection model is not static, so it is regularly updated with new data to adapt to evolving fraud patterns and to improve its accuracy. As part of these updates, we rigorously assess and address any emerging bias.