**Capstone Project**

**Signal that Speaks - Empowering Novice Investors with Data Driven Insights in ETF (Exchange Traded Funds) Investments**

**Girish Pandit**

**2024 - 2025**

# Problem statement

* What is the problem or the opportunity that the project is investigating?
* Why is this problem valuable to address?
* What is the current state (e.g. unsatisfied customers, lost revenue)?
* What is the desired state?
* Has this problem been addressed by other research projects? What were the outcomes?

# Industry/ domain

* What is the industry/ domain?

This project is within Financial Services industry with a focus on Exchange Traded Funds (ETF) and retail investing and specifically targets novice investors and aims to empower them with AI driven decision tools for ETF investments.

* What is the current state of this industry? (e.g. challenges from startups)

The investment management industry is undergoing significant transformation, especially in how retail investors interact with markets. Some notable developments and challenges:

Key trends:

* Rise of DIY investing : Platforms like Robinhood, eToro etc has enabled individuals to manage their portfolios directly.
* Explosion of ETFs: ETFs are rapidly growing due to their low cost, diversification, and ease of access.
* Digital transformation: AI and machine learning are being leveraged to offer insights, robo-advice, and predictive analytics.
* Regulatory focus: Increased scrutiny on suitability, data privacy, and transparency for retail-facing solutions.

Challenges:

* Information Overload: Retail investors often struggle with too much, often conflicting, data.
* Lack of personalization: Most platforms provide static or generic advice, not contextual insights.
* Fintech Disruption: Startups are challenging traditional players with better UX, faster innovation cycles, and data-driven advisory tools.
* Volatility & Uncertainty: Events like COVID-19 and inflation spikes make markets unpredictable, leaving investors anxious and reactive.

* What is the overall industry value-chain?
* What are the key concepts in the industry?
* Is the project relevant to other industries?

# Stakeholders

* Who are the stakeholders? (be as specific as possible)
* Why do they care about this problem?
* What are the stakeholders’ expectations?

# Business question

* What is the main business question that needs to be answered?
* What is the business value of answering this question? (quantify value and make necessary assumptions)
* What is the required accuracy? What are the implications of false positives or false negatives?

# Data question

* What is the data question that needs to be answered?
* What is the data required to answer the question?

# Data

* Where was the data sourced?
* What is the volume and attributes of the data?
* How reliable is the data?
* What is the quality of the raw data?
* How was this data generated?
* Is this data available on an ongoing basis?

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?
* What are the highlights of the Exploratory Data Analysis (EDA)?
* Is the pipeline reusable? (for example, to process future data?)
* What are the intermediary data structures used (if any)?

## Modelling

* What are the main features used?
* Did you find any interesting interactions between features?
* Is there a subset of features that would get a significant portion of your final performance? Which features?
* How did you select features?
* What feature engineering techniques are used?
* What are the models used?
* How long does it take to train your model?
* What are the tools used? (cloud platform, for example)
* What are the model performance metrics?
* Which model was selected?

## Outcomes

* What are the main findings and conclusions of the data science process?

## Implementation

* What are the considerations for implementing the model in production?

# Data answer

* Was the data question answered satisfactorily?
* What is the confidence level in the data answer?

# Business answer

* Was the business question answered satisfactorily?
* What is the confidence level in the business answer?

# Response to stakeholders

* What are the overall messages and recommendations to the stakeholders?

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?

# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)
* What are the resources used in the project? (libraries, algorithms, etc)