Chris Toomey, Jenice An, and Tongyu Guo
ADAN 8888 - Applied Analytics Project - Spring 2026
Module 1 Assignment - Project Report

HelpHerInvest: A machine learning-based investment recommendation system that aligns user interests and risk preferences with benchmark stock evaluation.

Problem Statement

Despite the growing availability of tools designed to support Americans with investing, a significant portion of the population continue to struggle with taking the initial step into the stock market. In a 2022 survey of 2,028 U.S. adults conducted by Stash, a personal finance app, 90% of respondents reported an interest in building wealth, yet almost half indicated that they do not know where to start. Additionally, the survey found that while the majority of respondents had basic financial accounts such as checking or savings, only 31% indicated owning a non-retirement investment account, highlighting a gap between financial intent and active participation in investing outside of employer sponsored plans (Stash, 2022). Beyond the findings of the Stash survey, broader national data reinforce this participation. Gallup (2025) notes that approximately 37% of U.S. adults in 2025 own no stocks at all, either directly or indirectly - a figure that has remained relatively consistent in recent years despite the expansion of online brokerage platforms and investment tools. This consistency suggests that increased access alone hasn't been sufficient to overcome the barriers many individuals face when attempting to enter the stock market.

Existing beginner-focused investment tools, such as Acorns, typically address this problem through full automation. It recommends and manages diversified portfolios of exchange-traded funds rather than guiding users through the selection of individual stocks or investment strategies (Unbiased, n.d.). While this approach simplifies the process for beginning investors, it offers limited insight into how specific investment choices are made or how users' personal interests relate to the specific investment in their portfolios. At the same time, many young and newer investors report using social media as a source of investment information. The SEC's Investor Advisory Committee cites findings that 60% of investors under the age 35 get investment information from social media, even as the SEC has issued multiple alerts warning that social media based stock tips can be misleading and are sometimes used to facilitate fraud (U.S. Securities and Exchange Commission [SEC], 2024).

This project seeks to address the gap between automated investing platforms and unstructured social advice by using machine learning to generate a recommendation system that translates individual preferences and risk profiles into personalized stock recommendations, with the goal of reducing knowledge barriers rather than promoting automated investment pathways.

Articulation of value

User value

For individuals who are interested in investing but unsure how to begin, this project offers a structured way to move from broad interest to actionable portfolio construction. The project's recommendation system narrows the investment universe by identifying companies that align with a user's stated interests and risk tolerance, reducing decision paralysis by limiting the number of options a user must evaluate.

Within this constrained set of interest-aligned stocks, the model applies machine learning methods to evaluate relative performance potential against relevant market benchmarks. As a result, users are presented with a curated portfolio of stocks that not only reflect their preferences but are also supported by data-driven indicators. This approach allows users to begin investing without requiring deep financial expertise, while still maintaining transparency around why certain stocks are included and others are not.

Market and Social Value

This project addresses a persistent participation gap in U.S. equity markets, where access to investing tools has increased but engagement has remained uneven. By focusing on structured decision support rather than automation or informal social advice, the project aligns with broader efforts to promote more informed and intentional participation in personal finance.

The framework emphasizes learning and transparency, which may help normalize investing and encourage a more sustainable approach to long-term wealth building.

Potential Target Market

Percentage of adults who do not own stocks = 37%[1]

Non-investors citing lack of knowledge or uncertainty as a barrier = 40%[2]

Percentage of adults willing to use FinTech app = 46%[3]

Conservative adoption rate = 1% - 10%

US adult population[4] j = **266,978,268**

Adults who do not own stocks 0.37 * 266,978,268 = **98,781,959**

Non-investors with lack of knowledge / uncertainty as a barrier: 0.40 * 98,781,959 = **39,512,783**

Willing to use fintech app: 0.46 * 39,512,783 = **18,175,880**

1% Adoption Rate = 18,175,880 * 0.01 = 181,759

10% Adoption Rate = 18,175,880 * 0.10 = 1,817,588

Target Audience Range = **181,759 - 1,817,588**

Potential Economic Value

Adults without stock ownership x % lack of knowledge/uncertainty = Individuals interested but unsure how to invest = our target market

Target market x conservative adoption rate = **181,759 - 1,817,588 users**

	1% Adoption Rate	10% Adoption Rate
$5 per month - $60 per year	$1,526,774	$15,267,739

[1] Gallup. (2025). What percentage of Americans own stock? https://news.gallup.com/poll/266807/percentage-americans-owns-stock.aspx

[2] World Economic Forum. (2022). New study finds financial education gaps are primary barrier to retail investing in capital markets. https://www.weforum.org/press/2022/08/new-study-finds-financial-education-gaps-are-primary-barrier-to-retail-investing-in-capital-markets/

[3] Cornerstone Advisors. (2019). Fintech adoption in the U.S. https://www.crnrstone.com/hubfs/19-0130_Q2_Fintech-Adoption-in-the-US.pdf

[4] U.S. Census Bureau. (2024). National population totals and components of change: 2020–2023. https://www.census.gov/data/tables/time-series/demo/popest/2020s-national-detail.html

13-Week Project Plan

Week 1 Identify the problem statement and dataset

1. Identify an opportunity that can be solved with machine learning
2. Research and determine dataset for the project
3. Articulate the value of solving the problem and develop a business case
4. Initiate project in GitHub and set up programming environment in JupyterHub

Week 2 Ingest and explore the dataset

1. Ingest data into JupyterHub and establish connection with VisualStudio
2. Identify what we are predicting for text classification model and multinomial classification model (i.e.,dependent variable)
3. Identify what we are using as features to make predictions for both models (independent variables)
4. Explore the variables in the dataset and brainstorm feature engineering

Week 3 Perform exploratory data analysis

1. Split our dataset into training, validation and test datasets
2. Perform EDA on our datasets
3. Identify problems, opportunities and pre-processing needed

Week 4 Make data model ready

1. Make data model ready by preprocessing our data based on findings from Week 3 and 4
2. Test feature engineering and create financial metrics
3. Brainstorm inputs to filter down stock selection (consumer interests, risk tolerance, etc.)
4. Perform preprocessing across training, validation and test datasets

Week 5 Engineer features

1. Create and engineer new features
2. Augment our dataset with new data
3. Reduce dimensions of our dataset

4. Create final training, validation and test datasets that will be used in modeling, evaluation and testing

Week 6 Develop 1st baseline modeling approach(es)

1. Build a simple set of classification models to solve our problem and dataset
2. Tune hyper-parameters of the model
3. Test out NLP approaches for retrieving user inputs for recommendation model
4. Select model evaluation metrics, evaluate variations and pick the winning model

Week 7 Develop 2nd complex modeling approach(es)

1. Build a complex set of models appropriate for solving our problem and dataset
2. Tune hyper-parameters of the model
3. Develop recommendation model to retrieve user inputs (possibly NLP approach)
4. Select model evaluation metrics,evaluate variations and pick the winning model

Week 8 Develop 3rd advanced modeling approach(es)

1. Build another set of models appropriate for our problem and dataset
2. Tune hyper-parameters of the model
3. Finalize recommendation model to retrieve user inputs for classification model
4. Select model evaluation metrics, evaluate variations and pick the winning model

Week 9 Select the winning model

1. Evaluate the developed models
2. Select the best recommendation and classification models
3. Calculate performance on the test dataset
    a. Test dataset will reflect the most recent financial performance compared to 9 weeks ago

Week 10 Data Centric AI

1. Improve our model by improving the data

2. Enhance feature engineering of independent variables and potentially develop new features
3.

Week 11 Explain the model, analyze risk, bias and ethical considerations

1. Explain our model by understanding feature importance and prediction outcomes
2. Identify model risks
3. Identify and quantify bias in our input dataset and model output
4. Identify and measure bias

Week 12 Save and package our model for deployment. Build our model monitoring plan

1. Save and deploy our model
2. Build a monitoring and maintenance plan for the classification model
3. Begin creating presentation and summarizing progress

Week 13 Review a peer's work and provide feedback

1. Review a peer's end to end modeling work and related artifacts
2. Provide written feedback to a peer's work and artifacts
3. Finalize the presentation and summarizing the model

The Dataset and Modeling Approach

The dataset our team has developed was obtained from an API called *yfinance* and contains various information on companies that trade in the U.S. stock market. The dataset contains 10,284 observations with 224 columns, and includes both text and numeric data. The textual data includes information such as the stock symbol, company description, industry, sector, and the company officers names / titles. Whereas, the numerical data includes market capitalization, total revenue, gross profit, return on equity, forward PE ratio, etc. We believe the dataset provides enough textual and numeric data for our team to develop a stock portfolio recommendation system, which aims to construct a portfolio based on the users inputs and provides stocks that outperform benchmarks in the market (such as the S&P 500). Our project deals with elements of supervised and unsupervised learning tasks.

Text Classification Model

We plan to develop a natural language processing (NLP) classification model to construct a subset of stock symbols based on the user's inputs. This would involve elements of unsupervised learning - such as topic modeling - where the model is extracting themes from the text to compare to user inputs. The output would result in classifying stock symbols and clustering them based on patterns in the company description, sector, industry, etc.

Multinomial Classification Model

We plan to develop a multinomial classification model to predict which stocks will outperform a benchmark (i.e. S&P 500) within the subset produced by the text classification model. This would fall under supervised learning because it requires a labeled dataset for the training inputs to then produce an output class (i.e. quartiles to categorize the stock rankings).

References

Gallup. (2025). What percentage of Americans own stock? https://news.gallup.com/poll/266807/percentage-americans-owns-stock.aspx

Stash. (2022). 90% of Americans want to invest, but almost half don't know where to start. https://www.stash.com/learn/90-of-americans-want-to-invest-but-almost-half-dont-know-where-to-start/

Unbiased. (n.d.). Acorns review. https://www.unbiased.com/discover/financial-advice/acorns-review

U.S. Securities and Exchange Commission. (2024). Recommendations of the Investor Advisory Committee on digital engagement practices and "finfluencers" (Investor Advisory Committee Report). https://www.sec.gov/files/approved-finfluencer-recommendations-20241210.pdf