

VIETNAM DATATHON 2023 - PROPOSAL

Team name: Bitminton || Team code: 40

1. Introduction

As you know, there is a lot of potential risk when a business owner needs to open a new store. Maybe something like not having enough customers, not having enough products to sell, having too many competitive stores,...

Business owners also want to know about potential customers, trending products in the future, and which month should they open their store to have the best starting.

Introducing our AI solution for businesses seeking strategic insights in the retail landscape. Leveraging artificial intelligence algorithms, our platform integrates predictive analytics to assist businesses in determining optimal store locations, identifying trending products on the horizon, and strategically pinpointing the most lucrative months for store openings. Showing most potential customers' data based on ages, incomes,... and history performance of stores in nearby locations. Our applications also show data in graphs to help a business owner know more about their data.

Gone are the days of relying solely on intuition and guesswork in the highly competitive world of retail. Our AI product empowers businesses to make informed, data-driven decisions, enhancing their competitive edge and positioning them for sustained success. Join us at the forefront of innovation as we redefine the way retailers navigate the complexities of the market, anticipate trends, and strategically plan their expansion efforts. Welcome to a new era of retail intelligence, where foresight meets profitability.

2. Problem Statement

There are a lot of myriad challenges in the retail industry. One of the most pressing issues is the strategic decision-making process involved in opening new stores and introducing products. Currently, businesses will face significant pain if they open new stores at the wrong location and don't have enough customers, they will waste a lot of resources. Four critical areas that businesses face are location selection for new stores, anticipation of trending products, the need for

efficient inventory management to serve their potential customers, and timing of store launches.

Optimal Store Location Selection:

- o In a lot of cases, businesses often struggle to identify the most suitable locations for their new stores, leading to suboptimal foot traffic and revenue generation.
- o Inaccurate location decisions result in wasted resources and capital, as companies may invest in areas with insufficient demand or fail to tap into lucrative markets.

Timing of Store Launches:

- o Determining the ideal month for store openings remains a challenge, leading to missed seasonal peaks and suboptimal initial store impact and profitability.
- o Inaccurate timing decisions can result in increased marketing costs, slower customer adoption, and reduced brand visibility.

Predicting Trending Products:

- o The ability to foresee upcoming trends in consumer preferences is a persistent challenge for retailers.
- o Without accurate predictions, businesses risk carrying outdated or unpopular inventory, leading to excess stock, markdowns, and missed revenue opportunities.
- o By knowing the upcoming trend, the business can increase its revenue and can avoid a lot of potential risks.

Efficient Inventory Management:

- o Businesses struggle with maintaining an optimal inventory level, often leading to excess stock, stockouts, and missed sales opportunities.
- o Inefficient inventory management contributes to increased holding costs, markdowns on unsold items, and potential customer dissatisfaction.

Competitor Positioning

Location Selection:

- o Traditional methods, GIS (Geographic Information System) tools, and basic demographic analyses are commonly employed.

Predicting Trending Products:

- o Many tools offer basic sales forecasting and trend analysis based on historical data.
- o Limited incorporation of machine learning for adaptive and real-time trend predictions.

Timing of Store Launches:

- o There aren't many tools that calculate the timing of Store Launches

Our AI-driven platform stands out in the market by offering a unified solution that addresses those retail challenges. We combine historical data with advanced algorithms to advise users about the trending products in the future, we also suggest store locations and opening times to help them gain the best revenue and increase their brand recognition.

3. Solution Overview

- Revenue level of each product will be classified based on consumers' behavior and standards of living. Then, within the classification, a sales forecast of that product will be generated.
- Due to the complexity of consumers' demand as well as fluctuation due to time, we cannot just simply predict the consumption rate of products simply from current trends or entirely by surveys (time and resource consuming).
- Because consumer needs are becoming increasingly common and complex, along with population displacement, it is impossible to simply predict consumer tastes through some basic observation of habits and lifestyles. Especially with new products, it's more unpredictable. Now, an AI model will analyze user tastes for items, combine with sales data when the business sells the product (if any) and give predictions about the potential of the product in the region.
- The product not only performs market research on sales data but also takes into account human factors (population, income, etc.), cultural factors (procedures, practices, etc.).

4. Methodologies

- Technology: the Random Forest algorithm and Scikit-learn library in Python.
- About our model:
 - + Collecting data:
 - Raw data of the industry in sales, volume, customers, etc. for the previous periods.
 - Separate the indices in accordance to province's range crawled from <https://www.gso.gov.vn/dan-so/>: GDP, population of individual age groups corresponds with category age_group within products.

- + Clean the data with available libraries:
 - Removing invalid data, *handling missing values*, outliers, etc.
 - Convert categorical variables into numerical ones using techniques like one-hot encoding.
 - Group the monthly sales data based on provinces to which they originate, treat the distribution channels belonging in the same province as a mutual distribution channel to collect data regarding the consumption behavior of that region.
 - Considering the culture reference of each locality, some *kinds* of product will be *manually* chosen to be biased.
- + Feature selection:
 - Use product attribute variables in combination with sales data to identify consumer behavior and purchasing power
 - Variables of products to consider:
 - age_group (related to population by age group): identify consumers' behavior by month
 - price_group (related to GDP): identify purchasing power
- + Decision trees and Random forest:
 - Number of variables for Decision Trees: tests all of the variables (from 1 to #vars), and then selects the value that gives the most accurate result.
 - Randomly choose variables: Bootstrap sampling
 - Tree growth: tree grows within a limitation of tree depth to avoid overfitting
 - Products classification:
 - Generate decision trees to classify products into 3 classes:
 - High revenue
 - Medium revenue
 - Low revenue
 - Aggregate results from decision trees to give the final result (based on the most voted result)
- + Give users a choice of products they want to sell, and then give them the corresponding sales forecasts for each month.
- + *Conclusion:* to what the model has been trained with, they can map population info, GDP,... to evaluate purchasing power and the consumption rate of a product in a province or city. Hence, it will have the ability to consider additional outlier variables (GDP and population info) when generating predictions in practice (testing, practical using,...).

5. Core functionality

- *Prerequisite:* to increase the accuracy of the results (serving for predicting consumer behavior), the user should provide the nearest sales data to retrain the model.
- The user will choose a location (province/city) in **Viet Nam** and a month to begin with in order to receive the forecasting result.
- The returned result will be Sales forecast per month in the selected region for the next 12 months, starting from the selected month.
- The product has two phases of operation:
 - + First, for every month, in the list of products that the company has distributed, based on the local sales situation during the same period and regional indicators such as population, GDP, age group, a list of products will be returned. This return list is a collection of **ALL PRODUCTS THAT THE INDUSTRY COVERS** classified into three categories (stated above)
 - + After returning 12 product lists, users have the option to choose the products they want from the list for selling. After that, the system will start calculating what will be the forecast revenue for **each product**.

6. Performance metrics

- Sales dataset is imbalanced: dataset will be classified into 2 unequal classes
 - + Products that have been sold before in the chosen locality (which means we have historical sales data of these products)
 - + Products that have not been sold before in the chosen locality (so that we will predict and suggest based on local indices such as population, GDP, groups of ages,...)
- Within the given dataset in the Online Test round which gives us the sales data of 19 months, we will split them into 2 parts using common split of 65-35 while also considering the aspect of a full year of data for objective analysis:
 - + The first part consists of 12 months in 2021 and is used to train our model.
 - + The second part is used to test the model. This will help test out the actual accuracy of the model created by comparing the results and seeing the fluctuation.
- Model Evaluation: The weight of *false negativity* and *false positivity* are equal such that we not only potentially avoid bad choices but also will

not accidentally skip good choices. Hence, we use the **F1-score** metric for evaluating performance.

- + Precision value: number of correctly labeled positive predictions out of all positive predictions.
- + Recall value: number of correctly labeled positive predictions out of positive instances
- + F1-score = harmonic mean of precision value and recall value

7. Timeline and Roadmap

Week	Date	Tasks and Milestones	Status
1	19/11 - 25/11	Kick-off, Data Review and Cleaning	In Progress
2	26/11 - 02/12	Algorithm Development, Analysis Data	Pending
3	03/12 - 09/12	Customer Data Analysis, GUI Development	Pending
4	10/12 - 17/12	Final Refinements, Documentation, Presentation	Pending

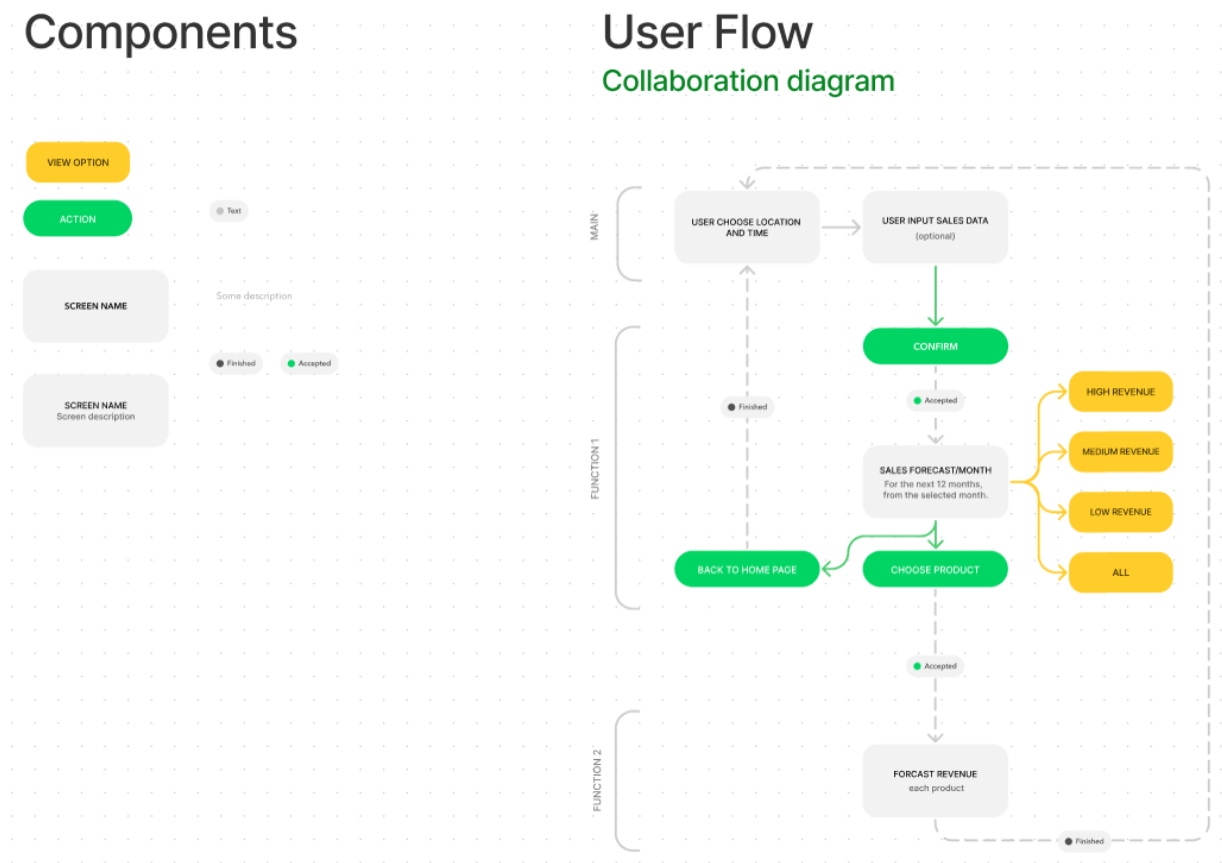
- Week 1: 19/11 -> 25/11
 - **Reading and Cleaning data:**
 - + Review and clean the **Inventory data files** for any missing and inconsistent information.
 - + Review **Master data files** to understand the structure and content.
 - **Researching:**
 - + Conduct research on best practices in retail analytics and AI-driven decision-making.
 - + Explore existing solutions in the market for similar retail challenges.
 - **Data Integration:**
 - + Begin integrating the cleaned inventory and master data files to create a unified dataset for analysis.

- Week 2: 26/11 -> 02/12
 - **Algorithm Development:**
 - + Start developing the AI algorithms for predictive analytics using the integrated dataset.
 - + Focus on store location optimization, trending product identification, and determining optimal opening months.
 - **Data Analysis and Visualization:**
 - + Analyze the integrated dataset to identify patterns and trends.
 - + Create initial visualizations (graphs, charts) to understand the data distribution.
 - **Refinement of AI Models:**
 - + Refine the AI models based on the initial analysis and feedback.
 - + Ensure that the models are robust and capable of providing meaningful insights.
- Week 3: 03/12 -> 09/12
 - **Customer Data Analysis:**
 - + Implement algorithms to analyze potential customer data based on age, income, and historical store performance.
 - + Incorporate customer insights into the overall decision-making process.
 - **Graphical User Interface (GUI):**
 - + Start developing a user-friendly GUI for business owners to interact with the AI solution.
 - + Include features for data visualization and insights presentation.
 - **Testing and Validation:**
 - + Conduct thorough testing of the AI solution.
 - + Validate the accuracy of predictions against historical data.
- Week 4: 10/12 -> 17/12
 - **Final Refinements:**
 - + Make final refinements to the AI models based on testing and validation results.
 - + Ensure the GUI is intuitive and provides the necessary information for decision-making.
 - **Documentation:**

- + Document the entire process, including data cleaning steps, algorithm development, and user guidelines.
- + Create a comprehensive user manual for business owners.
- **Presentation Preparation:**
 - + Prepare a detailed presentation outlining the AI solution, its capabilities, and the value it brings to business owners.
 - + Practice the presentation to ensure clarity and coherence.
- **Final Presentation:**
 - + Present the AI solution to judges, showcasing its features, benefits, and potential impact on strategic decision-making in retail.

8. User flow

The below diagram demonstrates the overview of user experience.



9. Limitations and future enhancements

- Limitations and Constraints of the MVP:
 - **Data Quality and Completeness:**
 - + The effectiveness of the AI solution is heavily dependent on the quality and completeness of the input data. Inaccurate or incomplete data could compromise the accuracy of predictions and insights.
 - **Limited Historical Data:**
 - + The MVP's predictive capabilities are constrained by the historical data available. A more extended historical dataset would contribute to more robust predictive models.
 - **Simplification of Customer Data Analysis:**
 - + The initial implementation focuses on basic customer demographics and historical store performance. The MVP may not capture nuanced customer behaviors or preferences, limiting the depth of customer insights.
 - **Assumption of Static Trends:**
 - + The MVP assumes that product trends and customer behaviors remain relatively static. In reality, market dynamics can shift rapidly, and future iterations may need to incorporate dynamic trend analysis.
 - **GUI Complexity:**
 - + While efforts have been made to create a user-friendly GUI, there might be challenges in accommodating all user preferences and levels of technical expertise. User feedback is crucial for refining the interface.
 - **Resource Constraints:**
 - + The MVP is developed under certain resource constraints, both in terms of time and technology. This may limit the complexity and scalability of the implemented features.

Potential Future Enhancements and Additional Features:

- **Enhanced Customer Segmentation:**
 - + Future iterations could include more sophisticated algorithms for customer segmentation, allowing businesses to tailor marketing

strategies to specific customer groups based on comprehensive behavior analysis.

- **Real-time Data Integration:**

- + Implementing real-time data integration capabilities would enable the system to adapt quickly to changing market conditions, ensuring that predictions and recommendations are based on the most recent information.

- **Advanced Machine Learning Models:**

- + Explore the integration of advanced machine learning models, such as deep learning or reinforcement learning, to further improve the accuracy and predictive power of the AI solution.

- **Dynamic Trend Analysis:**

- + Develop mechanisms for dynamic trend analysis that can adapt to rapid shifts in market trends, ensuring that businesses stay ahead of evolving consumer preferences.

- **Feedback Mechanism for Data Quality:**

- + Incorporate a feedback loop within the system to allow users to provide insights into data quality, helping to continuously refine and improve the accuracy of the AI models.

- **Integration with External Data Sources:**

- + Expand the system's capabilities by integrating external data sources, such as economic indicators, weather patterns, or social media trends, to provide a more holistic view of factors influencing retail performance.

- **Customizable Reporting and Dashboards:**

- + Provide users with the ability to customize reports and dashboards based on their specific needs, allowing for a more tailored and personalized user experience.

- **Scalability and Cloud Integration:**

- + Consider migrating the solution to a cloud-based infrastructure to enhance scalability, enabling businesses to analyze larger datasets and expanding the solution's capabilities as the business grows.

10. Conclusion

- In conclusion, the AI-driven retail solution we've proposed is poised to revolutionize how businesses approach the challenges of opening new stores. Our project, encompassing comprehensive data analysis, predictive analytics, and an intuitive graphical interface, is designed to empower business owners with actionable insights, facilitating informed decision-making. From meticulous data exploration and cleaning to algorithm development and a user-friendly GUI, our project systematically addresses key aspects of retail strategy, ensuring a holistic and data-driven approach.
- The value proposition of this initiative is profound. By leveraging advanced artificial intelligence algorithms, our solution not only optimizes store locations but also identifies trending products and strategically determines the most lucrative months for store openings. This transformative approach liberates businesses from reliance on intuition and guesswork, positioning them for sustained success in the highly competitive retail landscape.
- The potential impact of our AI solution is significant, offering business owners the ability to make informed decisions backed by data. This translates to improved operational efficiency, better resource allocation, and enhanced customer targeting. With the capacity to analyze potential customer data based on demographics and historical store performance, our solution opens avenues for personalized marketing strategies, increasing customer engagement and loyalty. In essence, our AI solution represents a powerful toolset for strategic planning, highlighting the transformative role of artificial intelligence in shaping the future of retail.