# Analytics Startup Plan

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | *Business Analytics Capstone* |
| **Requestor** | *Aisha Akter Anamika* |
| **Date of Request** | *14-7-25* |
| **Target Quarter for Delivery** | *August, 2025* |
| **Epic Link(s)** |  |
| **Business Impact** | *A predictive model built on classification analysis* |

## 1.0 Business Opportunity Brief

|  |  |
| --- | --- |
|  | Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment. |

Answer -

Breast cancer is one of the leading causes of cancer-related deaths among women in the U.S., and early detection remains critical for improving survival rates. This project focuses on developing a predictive model to classify breast tumors as benign or malignant using digitized cell-level data from fine needle aspirate (FNA) procedures, sourced from Wisconsin, USA. The goal of this analytics project is to identify key diagnostic indicators from digitized cellular data and to build a predictive model that classifies tumors as benign or malignant. This will support physicians with data-driven insights for faster and more accurate clinical decision-making. It will focus on feature analysis, pattern recognition, and model performance to improve early cancer detection and reduce diagnostic uncertainty.

In future it can be adapted globally, especially for individuals in underserved or resource-limited communities. The primary stakeholders include healthcare researchers, clinicians, and diagnostic developers, while secondary stakeholders include women undergoing cancer screenings and individuals interested in self-assessment tools. The goal is to build a reliable, interpretable machine learning model by August 2025 that promotes early awareness and detection. I chose this problem because I believe AI can make healthcare more inclusive and affordable. Without such tools, many people may continue to delay diagnosis due to cost or inaccessibility, leading to avoidable complications. Through this project, I hope to demonstrate how even simple data models can contribute to public health awareness and potentially save lives.

**The specific ask:**

Who – Women in the U.S.A particularly those at risk of or affected by breast cancer

What – A predictive mechanism that will be developed into a diagnostic tool to detect whether a cancer cell is benign or malignant.

Why – This modelling can make cancer diagnosis accessible and affordable on a preliminary level and can also raise awareness about the scope of AI in healthcare.

When – The model submission is on August 13, 2025.

So, what – Upon completion, the model will be open to public to provide feedback and work on. Additionally, in the future this can be considered as a medium of technology and can be adapted globally to have more refined versions.

## 1.1 Supporting Insights

|  |  |
| --- | --- |
|  | Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of the market, nationally and regionally? |

The landscape of breast cancer research and detection is increasingly leveraging data-driven predictive modeling, including machine learning (ML) and artificial intelligence (AI), to enhance diagnosis, prognosis, and treatment personalization. This approach moves beyond traditional methods to analyze vast datasets, identify complex patterns, and make more accurate predictions about disease progression, treatment response, and recurrence risk.

Key Competitors and Their Data-Driven Approaches:

Several companies are at the forefront of integrating predictive modeling into their breast cancer solutions:

**Exact Sciences Corporation**

Data-Driven Approach: Known for its Oncotype DX breast cancer test. This is a genomic test that uses gene expression data from a patient's tumor to predict the likelihood of breast cancer recurrence and whether chemotherapy will be beneficial.

Key Messages: Focuses on providing actionable insights through precision oncology to personalize treatment and improve patient outcomes.

Products & Services: Oncotype DX Breast Recurrence Score test.

Myriad Genetics, Inc.

Data-Driven Approach: Utilizes advanced genetic testing and data analytics for hereditary cancer risk assessment. Their tests analyze specific genes to predict an individual's risk of developing hereditary breast cancer, allowing for proactive screening and prevention strategies.

Key Messages: Aims to empower patients and providers with genetic insights to improve health outcomes throughout life.

Products & Services: BRACAnalysis CDx (for BRCA1/2 mutations), MyRisk Hereditary Cancer Panel.

**Illumina, Inc.**

Data-Driven Approach: A global leader in next-generation sequencing (NGS) technology. While not a direct diagnostic provider for breast cancer in the clinical sense, their technology is foundational for much of the data-driven research and diagnostics in oncology. They enable researchers and clinical labs to generate massive amounts of genomic data, which is then used by other companies and researchers to build predictive models for various cancer applications, including breast cancer.

Key Messages: "Unlocking the power of the genome to improve human health." They provide the tools that drive data-intensive genomic discoveries.

Products & Services: DNA sequencing instruments, reagents, and bioinformatics software.

**Tempus, Inc.**

Data-Driven Approach: A technology company that builds an operating system for precision medicine. They collect and analyze vast amounts of clinical and molecular data (genomic sequencing, real-world data) to help oncologists make data-driven decisions for cancer patients. They use AI and machine learning to find patterns and derive insights from complex datasets.

Key Messages: Accelerating precision medicine by making genomic data and AI-enabled insights accessible to physicians.

Products & Services: Comprehensive genomic profiling, bioinformatics support, and AI-powered insights for therapeutic guidance.

**Roche Diagnostics**

Data-Driven Approach: As a major player in diagnostics, Roche is increasingly incorporating digital pathology and AI into its workflows. This includes image analysis algorithms to help pathologists more quickly and accurately detect and characterize breast cancer cells, potentially leading to predictive insights on treatment response based on tissue morphology.

Key Messages: Focuses on delivering innovative diagnostic solutions that address patient needs and improve clinical outcomes through personalized healthcare.

Products & Services: Pathology solutions (including digital pathology systems), in-vitro diagnostics, personalized healthcare solutions.

**Market Share:**

It's challenging to provide a precise "market share" specifically for "data-driven predictive modeling" as it's an evolving technology integrated into various products rather than a standalone market segment. However:

Companies like Exact Sciences hold a significant market share in genomic profiling for breast cancer recurrence prediction, driven by the adoption of tests like Oncotype DX.

Illumina dominates the NGS technology market, which is the backbone for much of the genomic data used in predictive modeling, giving them an indirect but powerful influence.

The broader AI in healthcare market, including diagnostics, is experiencing rapid growth. North America leads this market, accounting for 45% of the global AI in healthcare market share, driven by significant investments and technological advancements

## 1.2 Project Gains

|  |  |
| --- | --- |
|  | *Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI’s.* |

To begin with, I would take a different approach by my very intention which is to empower individuals to better understand their own health. The goal is to make breast cancer symptom assessment more accessible and affordable by allowing users to compare their physical signs (like moles or lumps) with verified medical data.

For example, instead of paying hundreds for a basic visual exam, users could use advanced yet user-friendly technology such as image or pattern recognition tools to analyze patterns in their skin or tissue and gain a preliminary understanding of potential risks. This democratizes access to early detection, especially for those who can’t immediately consult a doctor.

Our focus is on affordability and accessibility, catering to people interested in AI and healthcare. By enabling self-assessment through data and technology, we give individuals the power to take control of their health while promoting awareness, education, and innovation in medical AI.

Our goal is to make the diagnostic tool as accurate and reliable as possible. However, like any medical technology, there are limitations, such as the risk of false positives or missing subtle indicators that a trained human might catch. But this challenge isn’t unique; it applies to all diagnostic tools. Machine learning models, like those evaluated with a confusion matrix, simply assess whether a prediction meets defined performance criteria. The real value lies in making healthcare more accessible by leveraging technology to provide timely, data-driven support, especially for those who might otherwise lack access to early diagnostic insights. This isn’t an attempt to replace radiologists, but rather to understand how they diagnose. What methods do they use and how we can simplify that process. The goal is to translate that expertise into accessible tools that empower patients to better understand their health and take informed action sooner.

I don’t aim to generate revenue from this project. My goal is to make the tool free and accessible at least during the trial phase. The key quality metrics I care about are the **accuracy of the predictive model** in detecting potential breast cancer cases and how **easy and intuitive** the tool is for users to navigate and understand.

If we don't do anything potentially, there will be bigger companies that are already working on AI to use in healthcare and to use AI to make more readily accessible cancer detection. It's just they're going to be making money off it for even simpler things because the public don't have enough knowledge on it or the data set on cancer patients are not readily available on open net so people from all over the world can contribute to it. It's just going to be another money-making method for big corporations. Instead of giving the public at least some opportunity or some access to contribute on this. and healthcare will keep being more expensive and inconvenient.

Attached below are two articles that show examples on how AI can transform healthcare

<https://www.rsna.org/news/2024/march/deep-learning-for-predicting-breast-cancer>

<https://www.theguardian.com/society/2025/jan/07/more-breast-cancer-cases-found-when-ai-used-in-screenings-study-finds>

I am working on a breast cancer dataset with **569 samples** of **breast cancer cell nuclei features and it was an open dataset. My time dedicated to this project would be 200 hours.**

## *Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.*

## 2.0 Analytics Objective

|  |  |
| --- | --- |
|  | List the key questions, assumptions and define the hypotheses. Often the delivery may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.  Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed. |

## Key Questions –

* What specific features are the strongest indicators of a malignant tumor?
* How accurate and reliable can a predictive model be in classifying tumors as benign or malignant?
* How can we simplify the diagnostic process for patients while maintaining clinical relevance?
* How secure is the data that will be fed into the system?
* What are the risks of false positives/negatives, and how should they be communicated to users?

Assumptions –

* The dataset used is representative of real-world breast cancer cases.
* Users of the tool will have basic digital literacy and access to smartphones or computers.
* Predictive modeling, even if not perfect, can still assist in early self-assessment and trigger timely medical consultation.
* Clinical diagnosis will still be required for confirmation. This tool acts as a support system, not a replacement.

## 2.1 Other related questions and Assumptions:

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis* |

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  | *What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.* |
|  | Model Performance Metrics   1. Accuracy/Precision (Positive Predictive Value)/Recall (Sensitivity / True Positive Rate)/F1-Score/ROC-AUC (Receiver Operating Characteristic – Area Under Curve)   User Metrics   1. Ease of Use / Time / Engagement Rate (for pilot phase) 2. User Confidence / Satisfaction Rate |
|  |  |

## 2.3 Methodology and Approach

|  |  |
| --- | --- |
|  | *Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.* |

**Type of Analysis:** Supervised machine learning- Binary classification analysis

**Methodology:**

**Exploratory Data Analysis (EDA):**

Assess data structure, distribution, outliers, and class balance using summary stats and visualizations.

**Data Preprocessing:**

Clean the dataset by removing irrelevant features, encoding target labels, and scaling features where needed.

**Feature Engineering:**

Select key predictors based on correlation and importance; created interaction terms if useful.

**Model Selection:**

Built and tested two classification models: Logistic Regression (for interpretability) and Random Forest (for performance).

**Model Evaluation:**

Compared models using Recall, F1-score, ROC-AUC, and confusion matrix to prioritize minimizing false negatives.

May develop more models as necessary.

**Output:**A trained and tested classification model that can predict whether a tumor is malignant or benign with high recall and accuracy. Clear explanation of the project and its business impact with a well-documented python script.

## 3.0 Population, Variable Selection, considerations

|  |  |
| --- | --- |
|  | Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today. |

**Audience/population selection:** Breast Cancer Patients from Wisconsin

**Observation window:** Biopsy period

**Inclusions:** Features that contribute to cells being benign or malignant

**Exclusions:** Any feature that does not contribute to identification

**Data Sources:** Open Dataset from Kaggle, specifically UCI Breast Cancer Wisconsin dataset

**Audience Level:** Individual

**Variable Selection:** Predictive and Correlated Features

**Derived Variables**: Depending on the state of variables, if any new variable is created it will be included. It’s yet to be determined.

**Assumptions and data limitations:** The two main limitations I found in my dataset are, first

the relatively small sample size, which may affect generalizability. Second, the dataset focuses

solely on cell-level features related to cancerous tissues, without including important

contextual factors such as the patient's clinical history, genetic background, or other medical

conditions that could influence breast cancer risk. These limitations reduce the scope and depth

of the analysis and model accuracy in real-world scenarios.

## 4.0 Dependencies and Risks

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Likelihood (based on historical data)** | **Delay (based on historical data)** | **Impact** |
| *Data quality*  Model performance  Overfitting | Medium  High  medium | Will be determined during execution of work | May affect Model accuracy  May require redesigning model  May need normalization of data |

## 5.0 Deliverable Timelines

|  |  |
| --- | --- |
|  | List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Item** | **Major Events / Milestones** | **Description** | **Scope** | **Days** | **Date** |
| 1. | Kick-off / Formal Request with Analytical Plan |  | Lock problem statement and dataset | Week 1 | 14/7/25 |
| 2. | Assessment of EDA |  | Inspect and visualize dataset | Week 1 | 20/7/25 |
| 3. | Feedback implementation on EDA |  | Review and implement feedback | Week 1 | 20/7/25 |
| 4. | Data Exploration & Analysis   * Issues * Solutions |  | Propose possible solution to fix the results derived from the dataset | Week 2 | 27/7/25 |
| 5. | Report development and Feedback implementation on python |  | Implement feedback from earlier EDA review | Week 2 | 27/7/25 |
| 6. | Data preparation and feature engineering evaluation |  | Carry out necessary steps to prepare the dataset and start building features | Week 2 | 27/7/25 |
| 7. | Model Exploration and Recommendation |  | Develop the most suitable model and explain why | Week 3 | 3/8/25 |
| 8. | Final Feedback |  | Review by Professor | Week 3 | 3/8/25 |
| 9. | Feedback implementation |  | Implement the feedback | Week 3 | 3/8/25 |
| 10. | Pilot |  | Present the model | Week 4 | 10/8/25 |
| 11. | Delivery & sign-off |  | Submission of model and report | Week 4 | 10/8/25 |