Predicting current and future war results from historic data (C8)

Karl Mumme, Taavi Raudkivi

Business Understanding Report

Background

Our project focuses on developing a predictive model to assist the military in strategic decision-making. The primary objective is to create a tool that enables the simulation of various war scenarios, helping military planners understand the potential outcomes based on different combinations of assets and factors.

Business goals

There are three main business goals:

1. Enhance Strategic Planning - provide a sophisticated tool for military strategists to simulate and analyze different war scenarios.
2. Optimize resource Allocation - assist in determining the most effective deployment of assets to maximize the chances of success in a conflict.
3. Improve Decision-Making - support military decision-makers with data-driven insights to make informed choices during planning and ongoing conflicts.

Business Success Criteria

Simulated scenario success rate - we will measure success by the model's ability to predict outcomes in simulated scenarios, with the goal of surpassing historical success rates.

Resource allocation efficiency - success will be demonstrated through a tangible reduction in resource allocation inefficiencies, leading to optimized deployment strategies.

Real-world conflict outcomes - ultimately, success will be evident in the real-world application of the model, where enhanced decision-making contributes to successful outcomes in military conflicts.

Inventory of Resources

To fuel our predictive model, we will tap into a rich array of datasets, like battles.csv, terrain.csv, weather.csv, front_widths.csv, etc. These datasets provide a comprehensive foundation for understanding historical battles, environmental conditions, and army assets. In terms of technology, our approach involves harnessing advanced data analytics and machine learning tools. These cutting-edge technologies will be instrumental in processing and analyzing the intricate relationships within the datasets.

Requirements, Assumptions, and Constraints

For data quality, we operate under the assumption of integrity and accuracy. Nevertheless, we remain vigilant and commit to identifying and addressing any data quality issues that may surface during the course of the project. Ensuring model interpretability is a crucial aspect of our strategy. We recognize the importance of making the model outputs easily understandable for military planners, facilitating seamless integration into their decision-making processes. With respect to the timeline, our commitment is to complete the project within the specified timeframe. This ensures the timely delivery of a valuable tool for military strategists.

Risks and Contingencies

Recognizing potential data limitations, we are proactive in addressing gaps or constraints within the historical data. Sensitivity analyses will be employed to mitigate the impact of these limitations on the model's predictive capabilities.

To manage risks associated with model complexity, we maintain a continuous feedback loop with end-users. Regular communication allows us to align the model with their understanding and expectations, ensuring its practical utility in real-world scenarios.

Costs and Benefits

The costs involved in this project encompass data acquisition, technology infrastructure, and personnel. These investments are justified by the anticipated benefits, including gains in strategic planning efficiency, resource optimization, and decision-making accuracy. The quantification of these benefits will serve as a tangible measure of the project's success.

Terminology

- Force Composition - the specific arrangement and types of military assets, including personnel, equipment, and technology, organized for a particular mission or operation.
- Theater of Operations - a geographical area, often large, where military operations are conducted, encompassing land, air, and sea components.
- Force Projection - the ability to rapidly deploy and sustain military forces over long distances to influence events in a given area.
- Force Multiplier - an asset or capability that increases the effectiveness of a military force beyond its inherent capabilities, often enhancing strategic advantage.
- Rules of Engagement (ROE) - definition: Directives issued by a military authority specifying the circumstances and limitations under which forces will engage in combat.
- Logistics Tail - definition: The logistical support system that ensures the timely and efficient flow of personnel, equipment, and supplies to and from the battlefield.
- C4ISR - Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance; a comprehensive system ensuring effective military decision-making.
- Battle Damage Assessment (BDA) - evaluation of the effectiveness of military strikes by assessing damage to enemy targets and the impact on overall mission objectives.

Data mining goals

Predictive Modeling for War Outcomes - develop a sophisticated predictive model leveraging historical data and relevant environmental variables to discern and forecast the outcomes of wars accurately.

Scenario Simulation Capabilities - enable the military to simulate diverse war scenarios, providing valuable insights into potential outcomes under varying conditions. This includes creating a tool that enhances strategic planning by considering a spectrum of potential scenarios.

Optimization of Resource Allocation - assist military planners in optimizing the allocation of assets by identifying the most effective combinations for different scenarios, ultimately reducing inefficiencies and enhancing overall strategic effectiveness.

Data mining success-criteria

High Prediction Accuracy - achieve a notable level of accuracy in predicting the outcomes of historical wars. The success of the model will be determined by its ability to consistently provide reliable predictions aligned with the actual historical outcomes.

Real-world Applicability and Impact - validate the practical utility of the model by applying it to current conflicts. Success will be measured by the model's ability to provide valuable insights that contribute to decision-making in contemporary military contexts, demonstrating its relevance and impact in real-world scenarios.

Data understanding

Gathering Data

Outline Data Requirements Our project requires data on historical battles and modern conflicts. For historical battles, we need data on battle name, date, location, strengths and losses on each side, victor, duration of the battle, and environmental and tactical environment descriptors. For modern conflicts, specifically the Russia-Ukraine war, we need data on personnel and equipment losses.

Verify Data Availability The required data is available on Kaggle. The data for the Russia-Ukraine war is available at https://www.kaggle.com/datasets/piterfm/2022-ukraine-russian-war/data and includes `russia_losses_personnel.csv`, `russia_losses_equipment.csv`, and `russia_losses_equipment_correction.csv`. The data for historical battles is available at https://www.kaggle.com/datasets/residentmario/database-of-battles and includes `battles.csv` and several other files providing additional context.

Define Selection Criteria We will select all data that provides insights into the factors contributing to the outcomes of battles and wars. This includes data on the strengths and losses on each side, the tactical environment, and the duration of the battle. For the Russia-Ukraine war, we will focus on data showing personnel and equipment losses. We will also consider any data corrections provided.

For our project, this is the first step in the CRISP-DM process. The next steps will involve describing, exploring, and verifying the quality of the data.

Describing Data

The data for historical battles, contained in `battles.csv`, includes information on over 600 battles fought between 1600 AD and 1973 AD. It provides details such as battle name, date, location, strengths and losses on each side, victor, duration of the battle, and environmental and tactical environment descriptors.

The data for the Russia-Ukraine war includes `russia_losses_personnel.csv` and `russia_losses_equipment.csv`, which contain information on personnel and equipment

losses during the war, respectively. The `russia_losses_equipment_correction.csv` file provides corrections to the equipment loss data.

These datasets provide a comprehensive view of both historical battles and modern conflicts, allowing for a detailed analysis of factors contributing to the outcomes of battles and wars.

Exploring Data

The data in the Historical Military Battles has to be cleaned in the sense that it has a lot of abbreviations and some of the column names don't make a lot of sense on their own. For instance, weather.csv only consists of on letter descriptions, which on their own say nothing. Fortunately, the dataset has a "datapackage.json." In it, there is a description for most of the columns of each table and even descriptions for some tables. To clean the data and give it meaning, we just have to use a short Python script.

The Russia Ukraine War dataset seems to be pretty cleaned up already and doesn't require a lot of work to understand. The column names are easy to understand, as is the data.

Verifying data quality

The quality of the data is overall satisfactory. There are no major concerns regarding the quality of the data. However, a minor concern is about the historical battle dataset. There are columns for time_min and time_max, and start_time_min and start_time_max, which are always the same. They ought to be consolidated into one column time and start_time accordingly. Another problem that doesn't really factor into the quality of the data but is important is that a portion of the data isn't really useful to us, such as commanders and front_width.

Project Plan

Overall Project Timeline:

- Day 1-2: Task 1 (Data Exploration and Preprocessing)
- Day 3: Task 2 (Feature Engineering and Selection)
- Day 4: Task 3 (Model Development and Training)
- Day 5: Task 4 (Model Interpretability and Validation)
- Day 6: Task 5 (Documentation and Presentation)
- Day 7: Finalize and Submit

Task 1: Data Exploration and Preprocessing

Objective: Quickly understand the structure and content of the datasets, address major data issues, and prepare data for analysis.

- Methods and Tools:
    - Python (Pandas, NumPy)
    - Rapid data visualization
- Comments:
    - Both team members actively participate in data cleaning and exploratory analysis.
- Time Allocation:
    - 4 hours

Task 2: Feature Engineering and Selection

Objective: Identify key features and perform basic feature engineering for model input.

- Methods and Tools:
    - Quick correlation analysis
    - Domain knowledge integration
- Comments:
    - Collaborative effort in feature selection and engineering.
- Time Allocation:
    - 5 hours

Task 3: Model Development and Training

Objective: Build a simplified predictive model using fast machine learning algorithms.

- Methods and Tools:
    - Scikit-learn for quick model prototyping
- Comments:
    - Both team members actively contribute to model development.
- Time Allocation:
    - 5 hours

Task 4: Model Interpretability and Validation

Objective: Ensure basic interpretability and validate the model's accuracy.

- Methods and Tools:
    - Basic interpretability techniques
    - Quick performance metrics
- Comments:
    - Collaborate closely on model interpretability and validation.
- Time Allocation:
    - 2 hours

Task 5: Documentation and Presentation

Objective: Quickly document the key steps and prepare a concise presentation.

- Methods and Tools:
    - Brief Jupyter Notebooks
    - Essential visualizations for presentation
- Comments:
    - Joint effort in preparing documentation and presentation.
- Time Allocation:
    - 2 hours