Vellore Institute of Technology, Chennai
Course: BCSE209L - Machine Learning

Predictive Modeling for Himalayan Expedition Success

Harnessing Machine Learning to Improve Safety, Decision-Making, and Risk Management

Instructor: Dr. Bhargavi R

Team Members:
- 23BAI1214 - Divyanshu Patel
- 23BAI1162 - Ayush Kumar Singh

Project Overview & Motivation

High-Risk Activity

Significant failure and fatality rates.

Financial Investment

Expeditions cost $35,000-$100,000+ per person.

Business Need

Outdoor companies need data-driven sponsorship decisions.

Safety Impact

Better prediction can improve expedition planning and safety.

Real-World Application

- Expedition companies can assess risk before investment.
- Insurance companies can calculate premiums more accurately.
- Climbers can make informed decisions.

Problem Statement

Himalayan expeditions involve extreme risks, with significant financial investments and life-threatening challenges. Companies, insurers, and climbers need reliable, data-driven tools to assess expedition risks more accurately. Current methods are often limited to historical trends and lack precise, predictive insights.

Objective

Develop a machine learning-based solution to predict the likelihood of expedition success defined as reaching the summit and returning safely achieving high accuracy and providing actionable insights for decision-making.

Key Questions to Address

What factors most strongly influence expedition success?
Which machine learning models are best suited for this complex problem?
How can predictive insights improve planning, safety, and investment decisions?
How can we ensure robust and generalizable predictions?

Dataset Description

Data Source

Kaggle Dataset: Himalayan Climbing Expeditions

- Source: **The Himalayan Database (archived records by Elizabeth Hawley, digitized by Richard Salisbury)**
- Geographic Coverage: Nepal Himalaya
- Time Period: 1905-2024

File	Records	Description
expeditions.csv	~11,000	Team-level expedition data
members.csv	~89,000	Individual climber records
peaks.csv	~481	Peak information and statistics

Key Features Available

- Climber Info: Age, sex, citizenship, experience level
- Expedition Details: Season, team size, hired staff, oxygen use
- Peak Data: Height, difficulty, first ascent history
- Outcomes: Success, death, injury, highest point reached

Existing Work Analysis

Current Research & Analysis

1. Himalayan Expeditions EDA (Kaggle)

- Focus: Comprehensive exploratory data analysis
- Key Insights: Peak difficulty vs success rates, seasonal patterns, age and experience factor analysis, success trends over decades.

2. A Century of Himalayan Expeditions, Visualized (Medium)

- Focus: Historical trend analysis and visualization
- Key Insights: Evolution of success rates, impact of modern equipment, geographical distribution, fatality rate trends.

Research Gap Identified

Limited Predictive Modeling: Existing work mainly focuses on describing past trends rather than predicting outcomes.

No Comprehensive ML Pipeline: Lack of an integrated, step-by-step approach for building and evaluating predictive models.

Business Application Gap: Absence of practical, deployable solutions for companies, insurers, or climbers to use in real-world decision-making.

Proposed Methodology & Improvements

Data Processing Pipeline

01

Data Integration

Join datasets, handle missing values, create unified dataset.

02

Feature Engineering

One-hot encoding, feature scaling, create derived features.

03

Feature Selection

Correlation analysis to remove redundant features, select relevant features.

Modeling Approaches

01

Multiple Algorithm Testing

Test various supervised learning algorithms, compare performance, select best model.

02

Model Evaluation

Cross-validation, multiple evaluation metrics (Accuracy, Precision, Recall, F1-Score), ROC curve and AUC score analysis.

Proposed ML Models & Selection

Logistic Regression (LR)

Why: Baseline linear model for binary classification.

Expected Benefit: Interpretable results and clear feature importance.

K-Nearest Neighbors (KNN)

Why: Instance-based learning for pattern recognition.

Expected Benefit: Captures local patterns in the data, effective for non-linear relationships.

Decision Tree (DT)

Why: Rule-based classification with clear decision paths.

Expected Benefit: Easy interpretation of decision rules and intrinsic feature selection.

Neural Network (Multilayer Perceptron)

Architecture: Feed-forward network with multiple hidden layers.

Expected Benefit: Ability to capture complex non-linear relationships in high-dimensional data.

Balanced Random Forest (BRF)

Why: Ensemble method specifically designed to handle class imbalance.

Expected Benefit: Improved accuracy and reduced overfitting, especially on imbalanced datasets.

Model Selection Strategy

- **Cross-Validation:** Stratified K-fold for robust and unbiased model evaluation.
- **Performance Metrics:** Comprehensive evaluation using Accuracy, Precision, Recall, F1-Score, and AUC-ROC curves.
- **Model Comparison:** Systematic comparison of all trained algorithms to select the best performing model based on defined metrics and business viability.

Technical Implementation Plan

Our project will leverage a robust stack of technologies, from programming languages and machine learning libraries to development environments and version control systems, ensuring a streamlined and efficient workflow.

Programming & ML Libraries

Python 3.8+ with leading libraries like scikit-learn, pandas, and numpy for powerful data manipulation and model building.

Development & Collaboration Tools

Jupyter Notebooks for interactive development, CSV files for data storage, and Git/GitHub for robust version control.

Data Processing & Visualization

Pandas and NumPy for efficient data handling, complemented by Matplotlib, Seaborn, and Plotly for insightful visualizations.

Expected Outcomes & Success

High Prediction Accuracy

Achieving **reliable and high predictive accuracy** for expedition success is a key outcome. This ensures trustworthy decision-making for sponsorship, planning, and risk assessment

Precision & Recall

Minimizing false positives and false negatives is crucial for precision and recall optimization. This enhances the reliability of actionable business insights.

Deliverables Overview

Key deliverables include trained ML models, feature importance analysis, and a prediction API. An interactive dashboard and a comprehensive technical report will also be provided.

Defined Success Criteria

Success will be measured by **robust model performance on unseen test data** and **identification of the most influential features**. Deployment-ready prediction systems form part of the final outcomes.

Demonstrating Value

Success will be measured by **robust model performance on unseen test data** and **identification of the most influential features**. Deployment-ready prediction systems form part of the final outcomes.

Project Timeline & Milestones

Our project will proceed through four distinct phases, each critical to the successful development and deployment of our predictive model.

Phase 1: Data Preparation

1

- Collect and clean data
- Conduct exploratory data analysis (EDA)
- Perform feature engineering and selection

This phase lays the foundation for the project by ensuring data quality and relevance for robust model training.

Phase 2: Model Development

2

- Implement baseline models
- Train advanced machine learning models
- Optimize hyperparameters for peak performance

This phase focuses on building and refining predictive models to achieve the highest possible accuracy and generalization.

Phase 3: Evaluation & Improvement

3

- Compare model performance using multiple metrics
- Select the best-performing model
- Implement ensemble methods for enhanced predictability

This phase ensures the chosen model is robust, reliable, and performs optimally across all critical evaluation metrics.

Phase 4: Documentation & Presentation

4

- Analyze and interpret all model results
- Prepare a comprehensive technical report
- Create the final presentation for stakeholders

This phase communicates our findings effectively and concludes the project with a comprehensive summary of our methodology and insights.

Challenges & Mitigation Strategies

Addressing the complexities of predicting expedition success.

Addressing Class Imbalance

Class imbalance arises due to a higher number of failed expeditions than successful ones, which can bias model training.

Mitigation: Utilize techniques like **SMOTE** (Synthetic Minority Over-sampling Technique), class weighting, and ensemble methods to balance the dataset effectively.

Managing Feature Complexity

The dataset contains numerous categorical variables (e.g., citizenship, peak name, season) that increase feature complexity.

Mitigation: Employ advanced encoding methods such as target encoding or one-hot encoding, and explore embedding layers for sophisticated processing of these features.

Handling Missing Data

Incomplete expedition records pose a significant challenge, as many entries have missing or inconsistent values.

Mitigation: Implement advanced imputation techniques (e.g., MICE, k-NN imputation) and robust feature engineering to address missing data effectively and minimize bias.

Preventing Overfitting

Building complex models on a potentially limited or noisy dataset risks overfitting, reducing generalization to new data.

Mitigation: Apply rigorous cross-validation strategies, regularization techniques (L1/L2), and early stopping to ensure model robustness and prevent overfitting.

Resources & References

Key materials and sources underpinning this project.

1 Key Dataset Resource

The primary dataset for analysis is available on Kaggle. It provides extensive data on Himalayan climbing expeditions, crucial for our predictive modeling.

2 Exploratory Data Analysis Reference

Further insights into the dataset's structure and trends can be found in an Exploratory Data Analysis (EDA) on Kaggle.

3 Visualization of Himalayan Expeditions

A compelling visual narrative of a century of Himalayan expeditions is detailed on Medium, offering graphical representations of historical data.

4 Elizabeth Hawley Archives

The foundational data originates from Elizabeth Hawley's extensive archives, providing unparalleled historical expedition records for the Nepal Himalaya region.

5 Scikit-learn Technical Reference

For our machine learning tasks, Scikit-learn is a vital library. Its comprehensive documentation is an indispensable technical reference.