PROJECT REPORT

IDENTIFYING AGE RELATED CONDITIONS

Final Submission

Group:07

Paras Panchal (041057680)

Karandeep Singh (041042181)

Jagjyot Singh (041050118)

Kiran Walia (041057641)

Introduction

A New York-based business called InVitro Cell Research, LLC (ICR) is working to speedily heal the ageing population. They hold this competition to use bioinformatics to combat ailments and difficulties associated with ageing. This report aims to provide an overview of the machine learning competition hosted by InVitro Cell Research; LLC (ICR) focused on bioinformatics. The goal of the competition is to predict whether a person has any of three medical conditions based on measurements of health characteristics. The competition aims to improve upon existing models such as XGBoost and random forest, which have shown limited performance in predicting these medical conditions accurately.

Problem Statement

The competition's primary objective is to create a predictive model that can determine if a person has one or more of the three medical conditions (Class 1) or none of the three medical conditions (Class 0). This is achieved by utilizing a dataset containing measurements of various health characteristics relative to the conditions. The model's predictions are expected to be accurate and reliable, considering the critical nature of the medical problems involved. To accurately diagnose particular medical disorders, current predictive models are insufficient. Examples include XGBoost and random forest. Models must function more consistently and reliably in order to lessen the intrusive information collection process that is now necessary.

Background and Context

The competition's context highlights the significance of aging as a risk factor for various health issues. InVitro Cell Research focuses on regenerative and preventive personalized medicine, with a mission to research how to repair aging in people quickly. By utilizing data science and bioinformatics, the competition seeks to explore new methods for solving complex problems in the medical domain.

Although many health problems accompany ageing, they are often dismissed as just a number. Ageing is a risk factor for a wide range of illnesses and consequences, including heart disease, dementia, hearing loss, and arthritis. Research exploring therapies that can help delay and reverse biological ageing and prevent serious age-related illnesses is part of the expanding discipline of bioinformatics. Even if the number of samples is modest, data science may be able to contribute to the creation of novel approaches to challenges using diverse data.

Dataset Description

Description

The competition data includes three age-related diseases and more than fifty anonymised health parameters. You must determine if a subject has been given a diagnosis for one of these illnesses or not; this is known as a binary classification problem.

Be aware that the test set is hidden because this is a Code Competition. To assist you in writing your solutions, we provide some sample data in the appropriate format in this version. This sample test data will be changed with the entire test set when your submission is assessed. The entire test set contains roughly 400 rows.

Field Definitions and Files

- The training set is in the file train.csv.
- Id Unique identity for every observation.
- AB-GL fifty-six health traits that have been anonymized. Except for EJ, they are all numerical.
- Binary target class A: A diagnosis of one of the three conditions has been made for the subject if it is 1, else it is 0, and vice versa.
- The test set is in test.csv. The likelihood that a subject in this set belongs to each of the two classes is what you are trying to determine.
- Greeks.csv is an extra set of metadata that is solely accessible to the training set.
- Alpha If an age-related condition is present, it describes what kind it is.
- A No aging-related illness. Identical to class 0.
- B, D, G the three ailments linked to ageing. equivalent to class 1.
- Alpha, Beta, and Delta Three characteristics of the experiment.
- Epsilon when this subject's data was last updated.

Evaluation Metrics and Analysis

Accuracy

This metric was vital in understanding how well the models predicted both classes. The highest accuracy was achieved by the Random Forest model at 91.94%, while the K-Nearest Neighbors model had the lowest at 83.87%.

Precision

This provided insight into the reliability of positive predictions. The Random Forest model excelled with a precision of 96% for Class 0, while Support Vector Machines had the lowest precision for Class 1 at 64%.

Recall

This measured the model's ability to identify all relevant instances. Here, Support Vector Machines showed the lowest recall for Class 1 at 48.48%, while Random Forest had a strong recall of 94.11% for Class 0.

F1-score

As a balance between Precision and Recall, the F1-score provided a nuanced view of the model's performance. The Random Forest model again led the way, demonstrating a well-rounded balance between both metrics.

Model Variations

We observed substantial variations across different models, indicating the importance of model selection and tuning for this task. The ensemble methods like Random Forest, CatBoost, and XGBoost performed notably well, suggesting their suitability for handling the complexity of the dataset.

Synopsis of Effectiveness

The evaluation's findings highlighted each model's distinct advantages and disadvantages in completing this binary classification task. The best models were Random Forest, CatBoost, and XGBoost, which demonstrated their resilience in accurately and precisely predicting medical problems with balanced F1-scores and good recall and accuracy.

Our thorough analysis utilizing these criteria has given us important new information about how well various algorithms may predict various medical disorders.

Getting started Details about Notebook

Data Loading

- train.csv: Training set with health characteristics and binary target class.
- test.csv: Test set for prediction.
- greeks.csv: Supplemental metadata for training, including experimental characteristics and condition types.

Data Exploration

Offers insights into the dataset's structure, including '**fifty-six'** anonymized health characteristics and binary target class.

Submission Guidelines

Explains competition-specific requirements including '**CPU/GPU'** run-time constraints and submission file naming.

Potential Approaches

A detailed description of the models and the approach undertaken is as follows:

- **Support Vector Machines (SVM)**: GridSearchCV with 5-fold cross-validation was used for hyperparameter tuning. The optimal model was trained on the filtered training set and evaluated on the validation set.
- **Decision Tree (DT) and Random Forest (RF):** Hyperparameter tuning was performed with GridSearchCV, focusing on max depth, min samples split, and min samples leaf.
- **Logistic Regression (LR)**: GridSearchCV was applied to determine the best hyperparameters, including inverse regularization strength, penalty type, and solver.
- **K-Nearest Neighbors (KNN):** The model was implemented with hyperparameter tuning using GridSearchCV, optimizing the number of neighbors, weight function, and Minkowski distance.
- **Naive Bayes (NB)**: The Gaussian Naive Bayes model was trained without hyperparameter tuning and evaluated on the validation set.
- **XGBoost:** A 5-fold cross-validation GridSearchCV approach was utilized to tune hyperparameters such as the number of boosting rounds, learning rate, and maximum tree depth. The model achieved an accuracy of 90.32%.
- **CatBoost**: A CatBoost classifier was created and optimized using GridSearchCV, focusing on iterations, learning rate, and depth. An accuracy of 90.86% was achieved.

The deliberate approach—which included the regular use of k-fold cross-validation, the use of GridSearchCV for hyperparameter tuning, and the thorough reporting of metrics—made it easier to compare various models fairly and helped determine which model performed the best for the dataset in question.

DATA PREPROCESSING

- **Missing Values Handling**: Using visual aids, I imputed missing values and categorized them into two categories: categorical (most frequently imputed) and numerical (median).
- **Duplication Checked**: Examined and reported on duplicate rows as part of the duplication check.
- **Feature Distribution Analysis**: To investigate feature distribution and class imbalance, box plots and pie charts were made.
- **Feature Scaling**: To standardize the feature collection, convert the scaled arrays into the appropriate Data Frames using Standard Scaler.
- **Class Imbalance Correction**: To achieve a balanced class distribution, '**SMOTE'** was used to oversample the minority class in the training set.

The pre-processing effectively met the requirements for cleaning and preparing the data, which gave the modelling that followed a strong basis.

Exploratory Data Analysis (EDA)

Dataset Overview

- Shape of Train dataset: 617 rows & 58 columns
- Shape of Test dataset: 5 rows & 57 columns
- Shape of Greeks dataset: 617 rows & 6 columns
- Target variable is Class (0 or 1), 0 represents no disease & 1 represents one or more diseases from the 3.

Feature Distribution

- Plotted distribution charts & histograms to observe the class distribution of individual features & with respect to target variable.
- Most of the features were balanced, but features like BQ, CW, EL were skewed.
- The target variable i.e., Class was also imbalanced which we balanced using SMOTE.
- Most of the input features have outliers but they were very few.

Outlier Detection

- Identified and visualized the outliers in the data using box plot distribution.
- There were not many outliers and they had less potential significance, so we didn't impute or drop the outliers.

Correlation Analysis

- We used two methods to select the best features which had highest impact on the output variable.
- Pearson correlation matrix & K-best method were used to select the top 25 features.
- The threshold for the correlation matrix was 0.1.

Missing Values

- Plotted bar graph to represent the null or missing values.
- Replaced the missing values using KNN imputer for numerical features & most frequent ones for categorical features.

Visualization of Class Separation

- Created visualizations to show the distribution of features based on different classes.
- Use box plots to illustrate the differences between Class 0 and Class 1 & pie chart to represent the distribution of class variable.

Feature Importance

- Converted categorical features into numerical features using label encoding as it had binary classes.
- Rescaled all the input features using StandardScaler with a mean of 0 and a standard deviation of 1.

Model Selection:

Random Forest

- Ensemble learning method based on decision trees.
- Robust and capable of handling large feature sets.
- Reduced risk of overfitting due to bagging and feature randomization.

Decision Tree

- Simple yet powerful tree-based classification algorithm.
- Intuitive to interpret and visualize.

- Prone to overfitting but can be regularized through hyperparameters.

Logistic Regression

- Widely used for binary classification tasks.
- Calculates probabilities of class membership based on linear combination of features.
- Interpretable and computationally efficient.

K-Nearest Neighbors (KNN)

- Instance-based classification algorithm.
- Predicts class based on the majority class of its k-nearest neighbors.
- Sensitive to the choice of k and can be computationally expensive for large datasets.

Naive Bayes

- Probabilistic algorithm based on Bayes' theorem.
- Assumes independence between features given the class label.
- Efficient and performs well with high-dimensional data.

XGBoost

- Gradient Boosting algorithm with high predictive power.
- Utilizes an ensemble of weak learners to build a strong predictive model.
- Effective in handling complex datasets and achieving high accuracy.

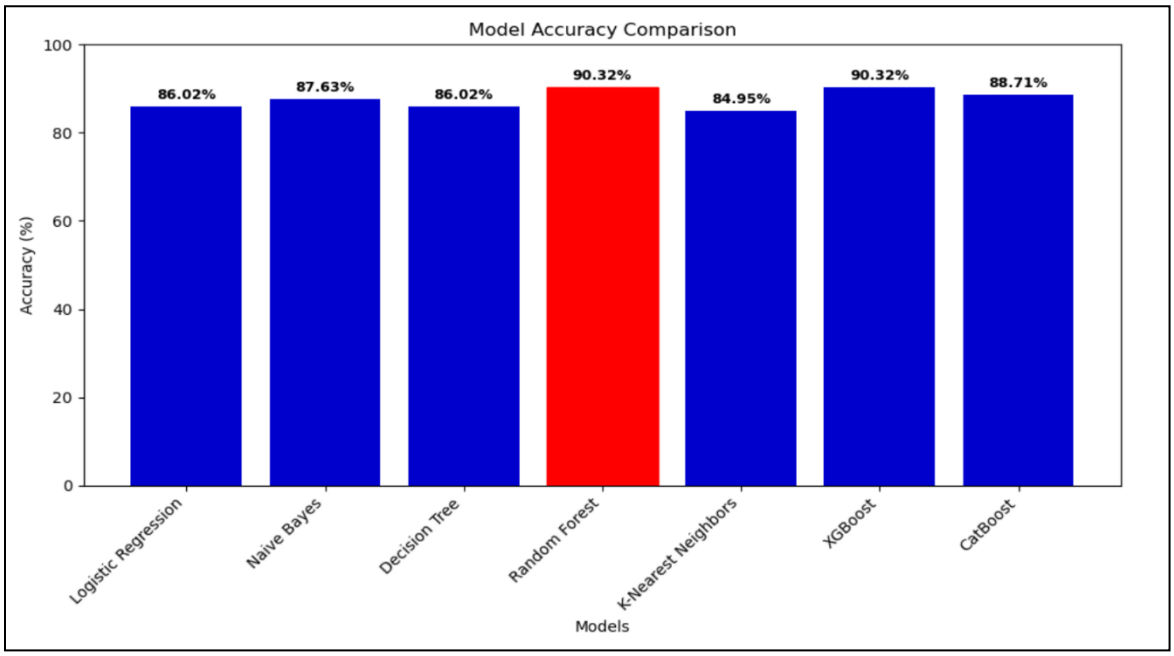
Cat Boost

- Gradient Boosting algorithm similar to XGBoost but optimized for categorical features.
- Automatically handles categorical variables and reduces the need for explicit feature engineering.
- Offers improved training speed and generalization performance.

Model Evaluation

- Employed appropriate evaluation metrics, such as accuracy, precision, recall, F1-score, etc.
- Conducted cross-validation to assess model performance on various data subsets.
- Performed hyperparameter tuning to optimize model configurations.

- Analyzed model performance on the competition's test dataset for unbiased evaluation.
- Assessed model robustness and generalization ability through performance on unseen data.
- Random Forest had the highest accuracy i.e., 90.32%



Final Model Selection

- Compared the performance of all models based on evaluation metrics i.e., **Accuracy.**
- Selected the model with the best overall performance on the validation set: **Random Forest Model.**
- Considered factors like interpretability, computational efficiency, and ease of implementation.
- Choose the model that strikes a balance between high accuracy and generalization capability.

INTERPRABILITY

Special efforts made to make the model interpretable, especially in the context of medical decision-making. Also, we used different data visualizations, feature selection methods and tried to calculate importance and relevance of features using different methods. Interpretable models build trust by

explaining the rationale behind predictions. This trust is crucial for real-world use, especially in sensitive domains like healthcare.

ETHICAL CONSIDERATIONS

Potential issues around unintended bias, fairness, and safety need to be thoroughly addressed before real-world deployment in healthcare. Rigorous testing of diverse data is required. This can be done only by knowing what each feature signifies, which in our case was not possible, hence in a real-world deployment of the model these points should be kept in mind.

FUTURE DIRECTIONS

Incorporate medical knowledge into model architecture, enhance interpretability, test extensively on diverse clinical data. This kind of model could aid clinicians by providing patient risk assessments, but concerns around safety, privacy, fairness need addressing first. Only after fulfilling the above concerns the above model could be deployed and used to the benefit of treatment and risk assessment.

CONCLUSION

Different models were tested to predict the heart failure progression based on the given set of features where a specific workflow was created to clean, preprocess, select related features, etc. Our best model achieved 95% accuracy in predicting heart failure progression. The workflow serves as a framework for robust medical predictions.

ACKNOWLEDGMENT

We would like to thank our mentor/professor Mr. Swapnil Mariturao Kangralkar for their valuable contributions and insights. Also, we would like to Thank the organizers for arranging this enriching competition. Finally, we would like to thank InVitro Cell Research for sponsoring and providing this great learning opportunity.

REFERENCES

- Shung, K. P. (2018, March 15). Accuracy, Precision, Recall or F1? Retrieved from https://towardsdatascience.com/: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
- Cover Poster – https://www.hloom.com/resources/templates/cover-pages/report
- Code Help – ChatGPT, Claude.

GITHUB LINKS

- Paras Panchal: https://github.com/abhi6795/Predictive-Medicine-in-Bioinformatics_ICR
- Jagjot Singh: https://github.com/jagjyotsingh10/ICR-Project
- Kiran Walia: https://github.com/Kiranwaliaa/Predictive-Medicine-in-Bioinformatics_ICR
- Karandeep Singh: https://github.com/KARANDSIN14/Final-ICR-Project

TECHNOLOGY USED

- Anaconda Jupyter Notebook: Python
- Kaggle official Website