EXECUTIVE SUMMARY

OF

STATE FARM CLASSIFICATION EXERCISE 40

BY FRANCK OLILO

The exercise 40 consist a classification problem given a train and test data , 40000 rows and 10000 rows respectively . The goal of the assessment was to clean and prepare the data, build two models and compare them.

The model I decided two build are the Logistic Regression and Gradient Boost machine Xgboost, In the followings lines, I will go over the approach use by each model and described their strength and weakness and will end of the paper by supporting which of the two models is better.

Approach

Logistic Regression and XgBoost are two popular machine learning models for classifications problems .Logistic regression is a model that assume a linear relationship between the predictors and response variable and require the response variable to be a binary (0 and 1) , It is a simple algorithm that is easy to implement , understand and conclude insights from. The approach of LR is to prepare the data , handle missing value , check the logistic regression assumptions such as no multicollinearity , relationship between response and predictors variable etc. after the model definition of the logit function , it use the maximum likelihood estimation to estimate the model parameter

On the other side , Xgboost is a mixed of regression and classification that use a gradient boosting algorithm for an optimal performance. It start with a single decision tree, compute the residual errors between the predicted value and actual values and fit a new decision tree on the residual errors abd can repeat the process until there is no significant improvement

Strengths and weaknesses

While those two models seem perfect as their strength, they also have weakness, in the table below I list some of their strength and weakness.

	Strength	Weakness
Logistic Regression	- Simple and easy to implement, making it a good baseline model. - It can handle both categorical and continuous independent variables. - It can provide interpretable results, allowing for a better understanding of the impact of the input features. - It can handle missing values through imputation.	- It assumes that the relationship between the input features and the output variable is linear, which may not always be the case. - It is sensitive to outliers and multicollinearity, which may affect the model's accuracy. - It is not well-suited for modeling non-linear decision boundaries.
XgBoost	- includes several regularization techniques like L1 and L2 regularization to prevent overfitting and improve the generalization performance of the model. - It is known for its high predictive power and is known to outperform wide range of machine learning problems.	- It can be difficult to interpret and explain - It has many hyperparameters that need to be tuned to achieve the best performance

Preference and Supporting reasons

While working of the data, train, and test data. I figure out after finding the problem of imbalanced data , that XgBoost will be the best choice for our classification .

Here as my supporting reasons :

- Xgboost is design to handle flarge and complex data with little information about the feature which is the case in exercise 40 we did not have any information about the attributes in the data .
- Xgboost is also designed to contain nonlinear relationship in contrast to logistic regression.
- it dealing internally with multicollinearity and overfitting problem by having the L1 and L2 regularization inside its algorithm process , which is not the case for logistic regression , that would have to be checked for overfitting and then decide whatever method is good to solve the overfitting method.