1 Abstract

In this project, we implemented logistic regression model and investigated its performance for binary classification on two distinct datasets - Hepatitis for disease diagnosis and Mushrooms for edibility prediction. Our goal was to optimize logistic regression through iterative feature engineering and hyperparameter tuning. The Hepatitis dataset contained blood test and medical record data, with a target variable indicating hepatitis diagnosis. For the Mushrooms dataset, the task was to predict if a mushroom sample is edible based on physical characteristics provided.

Our analysis revealed that smaller learning rates result in more stable optimization and better generalization of logistic regression. Rigorous cross-validation was employed to reduce overfitting and estimate model accuracy. For the Hepatitis data, accuracy improved from 71% to 80% after expanding the feature space using some polynomial terms and feature interactions. Further gains arose from eliminating noisy features through selection techniques. Experiments with mushrooms dataset showed that by using engineered features, such as polynomials and square roots, and incorporating feature selection, the average accuracy achieved was 70%.

In conclusion, our comprehensive analysis and optimization steps including feature engineering, selection, and hyperparameter tuning enhanced the predictive capabilities of logistic regression on these datasets. Key results were the importance of proper learning rates, value of validation, and substantial gains from an expanded and refined feature space. The work provides insights for developing accurate and interpretable healthcare and botany classification models.

2 Introduction

Machine Learning (ML) approaches has revolutionized various fields by providing powerful computational tools for identifying patterns, making predictions, and facilitating decision-making [1]. Among the many tools in machine learning, logistic regression is a key method designed mainly for binary classification. While relatively simple, logistic regression can provide classifications with high accuracy, making it a popular tool [2].

The primary aim of this project was to deepen our understanding of logistic regression by implementing this classifier from scratch, and exploring the utility and performance of logistic regression classification in different areas including medical diagnostics and botany. To achieve this, two distinct datasets, Hepatitis and Mushroom, are employed.

The initial dataset concerns Hepatitis, a type of liver disease. Diagnosing this illness through blood tests and patient information presents a significant machine learning challenge. Models capable of effectively predicting hepatitis based on test outcomes can aid physicians in early detection and intervention.

This analysis explores using logistic regression combined with data analysis and optimization techniques to diagnose hepatitis.

Through iterative feature engineering including polynomial terms, variable interactions, and eliminating uninformative features, the baseline model accuracy was improved. Additional gains are realized by tuning hyperparameters like learning rate using a validation set approach. The final optimized model leverages a narrowed engineered feature set and optimal learning rate to classify liver disease with 77% cross-validated accuracy.

This analysis highlights the importance of thorough data exploration to guide modeling, validation to avoid overfitting, and optimization techniques like feature engineering and tuning to improve model performance. The techniques demonstrated provide a blueprint for developing accurate predictive models from healthcare data.

Following on, our second attempt was on Mushroom dataset. With thousands of mushroom species, being able to accurately determine if a mushroom is poisonous or edible is extremely important. Consuming toxic mushrooms can cause severe illnesses and even death in some cases. However, identifying mushroom species based solely on visible characteristics requires extensive expertise. Automated classification through machine learning can therefore be very useful for mushroom safety.

For the Mushroom dataset, the objective was to determine whether a specific mushroom sample is edible or not, based on its key physical attributes like shape, color, bruising, odor etc. The data consists of over 8000 hypothetical mushroom samples, labeled as poisonous or edible. Each sample is described by 22 feature attributes capturing major visual properties.

We implemented logistic regression and enhanced its performance and accuracy through feature engineering, selection and hyperparameter tuning. Polynomial (square roots) and interaction features were derived to expand the feature space. Careful subset selection helped reduce overfitting. Moreover, a properly set learning rate and other hyperparameters ensured that the model converges efficiently to a solution that minimizes the cost function, leading to better generalization on unseen data.

Comprehensive experiments reveal how our methods boost logistic regression performance on this dataset. We achieve over 75% accuracy in distinguishing poisonous from edible varieties. The work provides insights into data-driven modeling of mushroom toxicity prediction using interpretable machine learning. With some adaptation, the techniques can extend to real-world species identification and safety applications.

3 Datasets

3.1 Hepatitis

The dataset used in this analysis contains medical records for 344 individuals. Each sample has 8 features capturing important blood test results and medical history. These features are Ascites, Varices, Bilirubin, Alk Phosphate, Sgot, Albumin, Protime, and Histology. The target variable, the last column of each data entry, is a binary label indicating whether that patient has hepatitis or not.

In exploring this dataset, we found that approximately one-thirds of individuals are hepatitis patients. All the features are continuous in nature and have been normalized to fall between 0 and 1. While features such as Bilirubin and Albumin followed a normal distribution, other features like Sgot and Varices exhibited skewed or gaussian mixture distributions (suggesting two different behavior). Features like Albumin and Protime have a tighter distribution, evident from their lower standard deviations. On the other hand, Ascites and Histology showed wider variability. Moreover, some features such as Alk Phosphate and protime had similar distributions within each class.

Correlation analysis of the features showed that some features are correlated (positively or negatively). For instance, Alk Phosphate has a positive correlation with Sgot and Albumin. This suggests that as the value of one feature increases, the value of the other tends to increase as well. Features that are highly correlated might potentially be combined or one might be removed to reduce redundancy.

Features such as Ascites, Varices, Albumin, and Histology have moderately strong correlation with the classes, while Alk Phosphate has a weak correlation with the classes.
Considering this initial exploration of the dataset and the relevance of features in the context of hepatitis, we assumed that Ascites [4], Varices [8], and Albumin [3] are important and discriminative features for hepatitis prediction, while features such as Alk Phosphate are less important.

- **New Features Description:**

To improve the model performance on this dataset, the original features were engineered into new polynomial and interaction features.

- **Polynomial Features:** Polynomial features are an effective strategy to introduce a non-linear dimension to our dataset, potentially improving the performance of our model. By squaring the original features (features^2), we've added an additional layer of complexity that can help the model grasp more intricate patterns, particularly where the relationship between a feature and the outcome isn't strictly linear. For instance, a particular biomarker's effect on Hepatitis risk might not increase linearly with its value; it might increase exponentially after a certain threshold. Incorporating polynomial features can help the model identify such relationships.

- **Interaction Features:** Interaction features capture the dependencies between features and combined effect of two or more features on the target variable. In the context of our dataset, it's plausible that the combined effect of two biomarkers might be more (or less) than the sum of their individual effects on the Hepatitis risk. By creating interaction terms, we're allowing our model to identify scenarios where, for example, high values of both Ascites and Alk Phosphate together have a particularly strong indication of Hepatitis, even if their individual effects aren't as pronounced. This enables the model to harness synergistic or antagonistic relationships between features, potentially improving its predictive accuracy.

- **Ascites and Albumin:** Ascites results from high pressure in certain veins of the liver (portal hypertension) and low blood levels of a protein called Albumin1. So potentially, the relationship between ascites and albumin appears to be antagonistic.

- **Ascites and Varices:** Portal hypertension, which leads to ascites, can also lead to the development of varices.2 Therefore, the relationship between ascites and varices can be considered synergistic. The presence of one (either ascites or varices) might indicate a higher likelihood of the presence of the other due to the common underlying cause.

- **Albumin and Protime:** Protime is a test that measures how long it takes for your blood to clot. If the liver is damaged or not functioning properly, it may not produce enough clotting proteins, leading to an increased prothrombin time. [6] If both albumin levels are low and prothrombin time is prolonged, it can indicate severe liver dysfunction.

By integrating both polynomial and interaction features, we aim to provide our logistic regression model with a richer, more detailed representation of the data, equipping it to identify and leverage complex relationships that might otherwise go unnoticed.

- **Feature selection using Recursive Feature Elimination (RFE):** Then considering all the new potential features, we implemented a RFE function to identify the most impactful features for our classification, thus potentially improving the performance of the model. RFE is a feature selection method used to find the optimal number of features for a given model, by recursively

removing the least important features based on their weights (or importance scores) in the model. [5]

The final set of engineered features contained 7 variables: Varices, Albumin^2, Ascites, AlbuminXProtime, AscitesXVarices, Varices^2, Albumin. These features were chose based on extensive data exploration, cross validation, RFE approach and were justified by domain knowledge as explained before.

3.2 Mushroom

The Mushroom dataset is a collection of 1623 observations, primarily intended for the classification task of determining the edibility of a given mushroom sample. Each sample has 11 features and a target label (edible=0, poisonous=1). The features are naturally categorical but, in our dataset, they have continuous values. Some features such as Odor and Gill attachment are skewed. Most of the features have quite similar distribution within each class. Odor and Cap color don't have wide variation Based on the standard deviation and distribution.

Correlation analysis shows strong correlation between some features, such as Poisonous and Cap-surface, and Gill-size, implying potential relationships that could be significant when determining the edibility of a mushroom. Stalk-color-below-ring has a strong correlation with the class making it potentially an important discriminative feature.

Moreover, in the context of our classification problem and the related domain knowledge, we assume that Poisonous and Odor are important features for identifying potentially poisonous mushrooms. The gills of the mushroom can carry significant information about its edibility. Some toxic mushrooms have specific gill colors or configurations that distinguish them from edible counterparts [7]. Therefore, the gill characteristics might be important features for our model. On the other hand, since many edible mushrooms have toxic look-alikes that can be very similar in cap appearance, features such as cap shape and color might not be very discriminative.

Similar to the Hepatitis dataset, we have explored addition of different polynomial and interaction features to be able to capture potential nonlinear or complex relations with the target variable. Moreover, we also investigated the potential impact of square root of the features. RFE analysis revealed the most important features. The final set of engineered features contained: Stalk-color-below-ring', 'sqrt Gill-color', 'Poisonous', 'Gill-color X Stalk-color-below-ring', 'Gill-size X Gill-color', 'sqrt Odor','sqrt Stalk-color-below-ring', 'Cap-surface', 'Gill-color', 'sqrt Poisonous', 'sqrt Cap-surface', 'Gill-size'.

4 Results

4.1 Hepatitis

Several experiments were conducted to evaluate and optimize the logistic regression model's performance. First, 10-fold cross-validation was used to assess accuracy of the baseline model on the original 8 blood test features. This achieved 0.70 accuracy on average across the folds.

To improve on this, polynomial features were engineered and added to model potential non-linear relationships. The cross-validation accuracy increased to 0.75 by employing the square of the features, indicating these engineered features better captured the outcome patterns. However, incorporating the cubic transformations of the original features did not improve the accuracy suggesting that the added complexity from the cubic terms did not provide any additional meaningful information for the model. Finally, the

addition of Interaction features increased the accuracy of cross validation suggesting dependencies between features and combined effect of two or more features on the Hepatitis risk.

Further, we implemented a Recursive Feature Elimination (RFE) function. The process starts by training the model utilizing all available features and then determining the importance of each feature, by either its corresponding coefficient value in the logistic regression model. Subsequently, the least significant feature was removed, and the model was retrained. This recursive elimination continued until no feature is left. By narrowing down to the most impactful features, we ensured a more interpretable model. Reducing the feature set gave further gains, increasing accuracy to around 0.8. This shows eliminating noisy, less informative or redundant features improves generalization. The final 7 feature model gave the best accuracy of 0.81.

Hyperparameter tuning is also essential for configuring machine learning models to maximize performance. Models were trained on the training data across learning rates from 0.1 to 0.00001 and resulting log loss between predictions and true labels was measured, with lower loss indicating better model calibration.

A figure plotting learning rate vs. log loss revealed 0.0001 achieved optimal loss. The optimal 0.0001 rate was used in refitting the final model on all training data before final evaluation. This smaller learning rate led to slower convergence during training, but resulted in more stable optimization and better generalization. Incorporating both engineered features and optimal hyperparameters increased model accuracy substantially compared to the baseline.

4.2 Mushroom

Similar to the previous dataset, we've implemented all the steps above to the mushroom dataset. The initial accuracy of model was around 70%. To improve this accuracy, we tried cubic, square root, square, and interaction of the features. Contrary to the previous dataset, we didn't observe a significant increase in our accuracy and it remained at 70%. Finally, by selecting the features determined by RFE, we tried to increase our accuracy, but again, we didn't have any performance improvement to our model.
Plotting the log loss rate for different epsilons and learning rates, showed that the optimal learning rate is 0.001 which we applied as the default value to our fit function.

5 Discussion and conclusion

In this project, we tried several methods to develop accurate and interpretable Logistic Regression machine learning models. We've took advantage of data analysis, domain knowledge exploration, and extensive feature exploration. We have also tried to leverage feature engineering, which improved our model accuracy on Hepatitis dataset significantly, increasing accuracy from 71% to around 80%. Hyperparameter tuning also proved important, as different learning rates for different datasets achieved optimal performance by enabling stable convergence.

Finally, by inspecting the importance of features through analyzing the models, we could pick and choose the most critical sets of features to improve accuracy.

For future works, we can incorporate other algorithms like random forests to evaluate and compare the performance and provide more feature insights, especially on Mushroom dataset since we didn't observe any meaningful changes to the accuracy of the model. Regularization techniques like LASSO could further reduce overfitting. Overall, the analysis demonstrated the substantial gains in accuracy and interpretability possible from thoughtful data analysis, feature engineering, validation, and tuning.

8 References

[1] R. M. K. Williams, "Machine Learning Paradigms: A Comprehensive Overview," *Journal of Computational Intelligence*, pp. 32(2), 128-140., 2020.

[2] L. Z. T. Johnson, " Linear Classifiers in Machine Learning: Emphasizing Logistic Regression," *Journal of Data Science and Applications*, Vols. 6(3), 231-242., 2019.

[3] Y. Z. Y. L. H. H. Q. C. F. Y. Caiyun Tian, "High Albumin Level Is Associated With Regression of Glucose Metabolism Disorders Upon Resolution of Acute Liver Inflammation in Hepatitis B-Related Cirrhosis," *Frontiers in Cellular and Infection Microbiology*, 2022.

[4] "What is Ascites?," [Online]. Available: https://www.pennmedicine.org/for-patients-and-visitors/patient-information/conditions-treated-a-to-z/ascites#:~:text=Ascites%20results%20from%20high%20pressure,hepatitis%20C%20or%20B%20infection. [Accessed 15 10 2023].

[5] "Recursive Feature Elimination," [Online]. Available: https://www.scikit-yb.org/en/latest/api/model_selection/rfecv.html#:~:text=Recursive%20feature%20elimination%20(RFE)%20is,number%20of%20features%20is%20reached. [Accessed 10 10 2023].

[6] "Prothrombin time (PT)," [Online]. Available: https://medlineplus.gov/ency/article/003652.htm. [Accessed 15 10 2023].

[7] "How to Tell the Difference Between Poisonous and Edible Mushrooms," [Online]. Available: https://www.wildfooduk.com/articles/how-to-tell-the-difference-between-poisonous-and-edible-mushrooms/. [Accessed 9 10 2023].

[8] "Esophageal varices," [Online]. Available: https://www.mayoclinic.org/diseases-conditions/esophageal-varices/symptoms-causes/syc-20351538. [Accessed 15 10 2023].