Binary classification using multivariate logistic regression for customer churn prediction

Overview of the data science pipeline and common libraries used

By Elliot Ledson

Overview of the solution

The code above is an implementation of a binary classification problem using multivariate logistic regression to predict customer churn in a bank. The data is first preprocessed by encoding categorical variables and scaling the data, and then used to train and evaluate the model. The code follows a typical data science pipeline and uses several common libraries such as numpy, pandas, scikit-learn, and matplotlib.

Details and next steps:

The implementation demonstrates flexibility in model hyperparameters, where a custom implementation of multivariate logistic regression is used to allow for control over the learning rate and number of iterations. The code also provides visualisations of the residuals and predicted values against the true values, which can aid in assessing the model's performance and identifying areas for improvement. Lastly, the bar plots of the churn rate for different columns provide insights into factors contributing to customer churn in the bank. To further improve the model's performance, implementing potential solutions such as SMOTE to balance the training data classes or feature engineering could be considered. Another thing is that I have a class make_classifications which is currently unutilised, which can generate random samples for each class, it would have been good to have concentrated on getting that implementation to a usable state.