EXAFLUENCE
Data Driven Influence

Platform-izing The ML Process

Malaya Rout
December 2023

A decade ago, we used to train a multiple linear regression model (MLR) with five independent variables and one continuous target variable without consuming much time and effort. More often than not, we used SAS or R. We called it a baseline (or base) model. With the advent of Generative AI and the disruption it is causing in various industries, a base model now implies one of the foundational LLMs by default. That is the paradigm shift we are experiencing in AIML today. This calls for completing our experiments using the "traditional" ML methods without spending much time.

The objective of platform-izing the ML process is exactly that. Standardization of the ML process bears fruit in terms of fewer human errors, faster experimentation, and saved money for the data science team in any organization. At Exafluence, we have built templates for data preparation, data exploration, and ML model experimentation. We also have templates for packaging and making the models accessible through APIs.

Q1.How would this platform be unique when other key AutoML players are in the market?

The uniqueness of the platform lies in the following.

- Automated and explainable data preparation with the right placement of interactivity as necessary
- Generation of a comprehensive data exploration report with insights using Generative AI

- Interpretation of various model evaluation metrics using Generative AI
- Selection of the best model with the right placement of interactivity as necessary
- A potential way of looking at the platform is as a Mixture of Experts (MOE) LLMs – One for data preparation, one for data exploration, and a third one for model training, evaluation, and finalization

Q2.What can we do to automate the data preparation phase in ML?

The data preparation module should produce the following automatically as soon as the raw dataset is provided as input.

- It should generate a summary statistics table with the number of continuous and categorical variables.
- There should be a missingness heatmap showing the spread of missing values. It should have a strategy for elimination versus imputation. In the case of imputation, for continuous variables, replace missing values with the average value. You might want to replace missing values in categorical variables with the most frequent class.
- Similarly, get the module to remove or impute outliers. Have a strategy for junk values similar to missing value.
- Regenerate summary statistics for the user to view the impact of changes in data preparation.

Q3.How do we automate the data exploration process in ML?

This module takes the prepared data set and creates a report of data exploration.

- In univariate analysis, generate a histogram and a boxplot for a continuous variable. Have a strategy for binning if necessary. In the case of a categorical variable, devise a frequency table of each class and a strategy for binning if necessary. Generate takeaways from the univariate analysis.
- In bivariate analysis, focus on all possible pairs. Generate scatter plots for continuous/continuous pairs. Generate boxplots having multiple boxes in one chart for continuous/categorical pairs. Do cross-tabulations and heatmaps for categorical/categorical pairs.

- The insight generation from the above charts can be done using Generative AI to make the report meaningful and appealing to the user.

Q4.Any Thoughts on Automating the Supervised Modelling Steps in ML?

Identify whether your problem statement can be solved by supervised learning.

- If yes, from the target variable, one can determine whether to use either a regression or a classification technique. Again, the target variable will tell us whether we deal with binomial or multinomial classification. Allow the user to input the train/test distribution numbers.
- Build and test the following models for regression technique. Linear Regression (pick up SLR/MLR based on the number of independent variables), Regression Decision Tree, Regression Random Forest. Create a model evaluation strategy. Display all diagnostics and comment on over-fitting. Plot bias/variance and display the optimum point. Create a comparison chart or table for all of the above regression techniques. Arrive at an overall diagnostic score. Recommend the data scientist to go with the best one. Combine train and test datasets and rebuild the model with the best technique recommended above. The idea behind using the test data along with the train data to train the final model is to ensure that we don't skip the test data unnecessarily from getting included in the final model created.
- Build and test the following models for binomial classification and apply the model evaluation strategy to all. Focus on Logistic Regression, Classification Decision Tree, Classification Random Forest, Naive Bayes, KNN, SVM, and Neural Networks. Device the comparison chart among all techniques. Recommend the best one.
- Build and test the following models for multinomial classification and apply the model evaluation strategy to all. Consider Multinomial Logistic Regression and Random Forest. Draw a comparison chart between the two. Pick the best one.

Q5.Can we automate clustering?

For the moment, let's keep hierarchical clustering aside and go ahead with K-means and K-prototype clustering. Take all numeric variables as input. Determine the optimum number of clusters using the elbow curve. Build K-means and K-Prototype clusters. Remember that the former technique uses only numeric variables, while the

latter can deal with a combination of numeric and categorical variables. For K Means, the BSS (Between Sum Square) should be as high as possible, and the WSS (Within Sum Square) should be as low as possible.

Q6.How About Association Mining and Time Series Modelling?

Use the Apriori algorithm. Give transactions data set as input. Create a diagnostics report using support, confidence, and lift. Choose the top rules. Generate insights using Generative AI from the selected rules.

For time series modelling, check for the stationarity of the series. Use the ARIMA modelling technique. Build models with varying values of p, d, and q. Recommend the best one.

Conclusion

Automation and abstraction make data preparation, data exploration, model training, and model evaluation easier and faster. However, with increasing levels of abstraction, the explainability of what happens under the hood decreases. Hence, special attention needs to be given to such tools to keep the data scientist aware and powerful by making the automated decisions explainable.

If you would like to learn more, write to us at marketing@exafluence.com

For interesting videos about our solutions subscribe to our YouTube channel- https://tinyurl.com/YouTubeExf

For regular updates on our solutions follow us on LinkedIn https://tinyurl.com/LinkedInExf