**Supplier Audit Results Prediction**

Data Science Internship Case Study

* Welcome

Hello everyone, my name is Juan Betancourt and I’m a master’s student applying for a data science internship position. In this presentation I will share my approach and results to the proposed task of predicting supplier audit results.

* Contents

These are the contents covered in this presentation. Let me begin with a brief introduction to the problem.

Introduction (3 mins)

* Motivation

Managing suppliers is a critical aspect of ensuring the operational efficiency of any manufacturing company. Challenges such as irregular lead times, backlogs, and incorrect orders can significantly disrupt manufacturing schedules, increase costs, and ultimately impact the service level provided to customers.

In the proposed context, our focus lies in the supplier audit processes designed to ensure reliability of suppliers. Specifically, selected suppliers undergo a thorough audit process before being fully entrusted with delivering parts for BMW production processes.

* Context

Given this context, let's delve into the problem definition. BMW relies on a group of suppliers,

each supplying a distinct set of derivatives for the production process. It's important to note that every supplier works with a specific subset of derivatives. The auditing process is designed to assess each supplier's performance concerning the derivatives they provide.

For a particular supplier and a particular derivative, the audit unfolds in three sequential stages: T1, T2, and T3.

At each stage, suppliers are assigned a color-coded qualification: green for a qualified status, yellow for a warning, and red for disqualification. The ultimate audit result is determined by the qualification assigned in the latest audit stage.

To illustrate this dynamic, let's consider an example where a supplier undergoes an audit for a given derivative. In the initial time step, the supplier might receive a green qualification in the first stage, T1. Consequently, the audit result at that time step would be green. Moving to the next time step, the supplier advances to T2 and is assigned a red qualification. The audit result, in this case, becomes red since T2 is the latest active stage. In the next time step, if the supplier attains a yellow qualification for T2, the audit result changes to yellow.

For the last two time-steps, the supplier progresses to T3, and the audit result reflects the qualification of this stage, given that it is the latest audit stage. Our primary focus is on predicting these audit results, anticipating whether there will be a sufficient number of qualified suppliers in the future.

Data Exploration (3,5 mins)

* Summary

To delve into this scenario, two datasets are at our disposal.

The first dataset comprises audit records for supplier-derivative pairs. Each observation encompasses the month of the audit, the derivative's region, the production line, qualifications for all three audit stages, and the overall audit result.

Key points to note about this dataset: it has over 7600 observations spanning 9 months, involving 818 suppliers and 75 derivatives. Suppliers exhibit varying entries, ranging from 1 to 21.

On the other hand, the second dataset encapsulates supplier performance logs. Each entry details wrong deliveries and backlogs for the last 3, 6, and 12 months, the LPKM score (a 5-star rating), and a binary indicator denoting past bad performance.

This dataset, with 5823 observations, one per supplier, provides a noteworthy insight—the majority of suppliers lack audit records.

Let’s gain a better understanding of both datasets.

* AH Distributions

Sarting with the Audit Historical Data, one notable observation is the dominance of derivatives from Europe, with South Africa contributing only a minimal proportion.

The second interesting insight emerges when we evaluate the results for each audit stage. A cursory glance reveals a clear imbalance in classes across all stages. Notably, the first stage, T1, lacks non-audited observations entirely. On the other end, the last audit stage, T3, exhibits a significant number of non-audited observations. T2 falls somewhere in between.

* AH Missing values

This dataset only has missing values in one feature, in 143 observations. I opted to remove these observations with missing values.

Nevertheless, there are more sophisticated methods for handling missing information, and I will discuss these approaches towards the end of the presentation.

* SP Distributions

Now, onto the second dataset. A glance at the x-axis scale reveals that the count of backlogs is significantly higher than that of wrong deliveries.

Additionally, the proportion of suppliers with a bad performance indicator is notably low, indicating a small number of potentially delicate suppliers.

* SP Correlogram

Given that most features are numeric, a correlogram proves beneficial in visualizing the relationships between them.

Firstly, we observe a positive correlation between the counts of wrong deliveries for 3, 6, and 12 months. The same holds true for backlogs.

As a final observation, there exists a negative correlation between the LPKM score and all other variables. This aligns with the logic that higher scores indicate better suppliers, less likely to incur in wrong deliveries and backlogs.

* SP Missing values

Concluding our data exploration, I applied one-hot encoding to the bad supplier indicator and removed the six observations containing missing values.

The same remark made earlier for handling missing data applies here as well.

Data Preprocessing and Feature Engineering (3,5 mins)

* First

The initial steps in the data preprocessing involved removing a feature with redundant date information and encoding all categorical features numerically.

* Overall approach

As previously mentioned, numerous suppliers have limited or no audit records. Consequently, the Machine Learning algorithm's features cannot rely on a complete audit record to predict the outcome. Considering this and aiming to predict results for supplier-derivative pairs, it is imperative to incorporate information from both entities.

* Feature selection

To begin, we can gather information about suppliers, including the LPKM score and the bad performance indicator. For derivatives, we can access details such as the derivative region and the production line.

Recalling the two datasets, there is additional information about audit results and the occurrences of wrong orders/backlogs over the past months.

* WO and Backlogs in the past 3 months

The next two features are founded in a primordial assumption, that the two data sets are chronologically linked.

This means that the 3, 6 and 12 months prior indicated in the Supplier Performance dataset are relative to the latest audit records.

To be more precise, the audit history dataset has observations between February and October of 2022. Therefore, the months prior will be relative to the month of November. Assuming the chronological link, two new features can be introduced: Wrong orders and backlogs in the past three months. All observations in August, September, and October would receive the WO and Backlogs of the 3 months prior. May, June, and July would receive the WO and Backlogs generated in those three months. This can easily be computed by subtracting the WO and Backlogs at 6 months minus the WO and backlogs at 3 months. This way, the features that had high correlation are encapsulated into only one.

* Previous Audit results

Additionally, it is just sound to think that the result for an audit stage is highly dependent on the previous result. Therefore, another feature I consider is the time elapsed since the last audit and the received qualification. This feature incorporates the sequential qualification scale.

Naturally, the first audit record for all supplier,derivative pairs are discarded as they have no ‘previous’ audit.

* Summary

To summarize, these are the features considered.

The features were chosen based on contextual logic. However, optimal methodologies can assess the importance of features and select them based on their ability to explain the response.

Model Selection and Prediction (3 mins)

We're dealing with a multiclass classification problem, aiming to predict which category an observation falls into—whether it's non-audited, red, yellow, or green.

The challenge intensifies because the ultimate goal is predicting the audit result, which hinges on the qualifications of all three stages (T1, T2, T3). Now, if we create a single classifier to predict the overall qualification, we might oversimplify the intricate dynamics unique to each audit stage. That's why I've chosen a more nuanced approach—predicting each audit stage separately.

This means deploying three distinct models, each dedicated to forecasting the outcome of T1, T2, and T3. Subsequently, these stage-specific predictions become inputs for another model, helping us deduce the comprehensive audit result. There are alternative strategies, such as a unified model considering all possible combinations of T1, T2, and T3. However, I'm inclined towards individual predictions for a more granular understanding of the process.

To predict each audit stage three algorithms were assessed and compared:

Random Forest:

Logistic Regression:

Support Vector Classifier:

* Evaluation metric

In assessing and contrasting the models, my primary metric was the mean accuracy derived from a 5-fold cross-validation.

Yet, considering the context and the business problem, another objective could be minimizing the count of suppliers erroneously qualified as green, a factor crucial for fortifying the robustness of the outcome.

* Best model

After implementing models for all the stages, I found that for the first two stages, T1 and T2, the best model was a Random Forest, and for the last audit stage T3, the best model was a Logistic Regression.

An important aspect to mention is that I implemented the models with the default parameters, chose the best one, and then calibrated the hyperparameters. However, the adequate order of things would be to compare all the algorithms when they are already tailored for the specific problem.

For both models, hyperparameter tuning was performed via a Grid Search. The hyperparameter tuning for the random forest encompassed the number of estimators, the max depth, and the minimum samples per split and leaf. The hyperparameter tuning of the logistic regression encompassed the c value and the penalty type.

* Audit results

In my methodology, there is an intermediate step missing: translating the individual audit stage results into an overall audit result. For this purpose, I employed a simple heuristic that checks the latest non-'non-audited' stage and selects that as the audit result. This yields an accuracy of almost 60% in the audit results. Nevertheless, the individual models maintained a high accuracy in predicting the qualification of specific stages.

This step has room for substantial improvement. By integrating another machine learning algorithm that takes as input the predictions for T1, T2, and T3, along with potentially additional features, the overall performance could see a significant boost. Moreover, the implementation of probabilistic models could be explored to better capture the dynamics of how audits are conducted, and qualifications evolve over time.

Business Recommendations (2 mins)

Model Maintenance (2 mins)

Further Development (2 mins)