Governance Monitoring and Validation

Introduction  
Maintaining model correctness and trustworthiness is critical in the field of predictive modeling. Effective validation and governance techniques are critical to maintaining model reliability. This research studies the dynamic balance between model validation and governance in the context of airline passenger satisfaction prediction. The research provides insights to stakeholders in the ever-changing predictive modeling ecosystem by investigating this interplay. The project not only underlines the necessity of validation and governance through this investigation, but it also gives solutions to mitigate risks and biases. The project contributes to informed and ethical decision-making by providing practitioners with tools to maintain reliability. Validation and governance are foundations of assurance in an era where predictive modeling changes outcomes, assuring long-term success.

Variable-level monitoring  
Accurate and dependable outcomes in predictive modeling necessitate continual vigilance. In a dynamic context, variable-level monitoring arises as a critical strategy for ensuring the consistency and relevance of input variables. Monitoring at the variable level entails examining the behavior and performance of individual input variables used in prediction models. This emphasizes the importance of continual assessment to detect potential anomalies such as outliers and properly address missing values using the median imputation technique. For the treatment of outliers, we use Cap and Floor technique where we capped and floored all our outliers which are beyond the upper bound and lower bound values respectively using Inter Quartile Range. This process results in a defined range, which subsequently becomes the acceptable range for the attributes. When dealing with missing or null values, a consistent approach is used—these values are replaced with the appropriate attribute's median. The median was purposefully chosen because, unlike the mean, it is a robust central tendency metric that displays robustness to extreme extremes.

Variable drift monitoring  
Variable drift is the term used to describe changes in the distribution or characteristics of input qualities over time. Identifying which variables are likely to drift and identifying their significance can have a major impact on model performance. Certain parameters, such as 'flight\_distance,' 'inflight\_wifi\_service,' and 'departure/arrival\_time\_convenient,' are especially vulnerable to variable drift due to changing customer preferences, external influences, or evolving industry norms. These characteristics influence passenger perceptions and are critical in predicting levels of satisfaction. Variable drift monitoring is especially important for passenger experience attributes such as 'food\_and\_drink', 'online\_boarding', 'seat\_comfort', and 'inflight\_entertainment'. Changes in these qualities can have a major impact on the model's predictive power as client expectations change.

Variable drift tolerance  
The drift tolerance for more important features i.e., ‘type\_of\_class’, ‘type\_of\_travel’, ‘inflight\_wifi\_services’ is set to be 6% of their mean values used in training, while drift tolerance for less important features i.e., ‘online\_boarding’ and ‘customer\_type’ is set to be 10% of their mean values. If variable drift occurs and exceeds the permitted range, various reasons will be investigated such as:- is the drift due to changes in the data collection method, is it due to the change in the data's underlying characteristics, etcetera? The impact of the drift then will be evaluated on the model’s performance using various techniques.

Model Monitoring, Health & Stability

Building a machine learning model is not an easy task. It is significantly more difficult to put a service into production. Even if you were successful in connecting all the pipelines, the story does not end there. Once the model is in use, we must immediately consider how to keep it running smoothly. Any disturbance in model performance results in actual business loss. We must ensure that the model performs as a machine learning system that we can rely on to make judgments (Dral, 2021).

A diagram of a service

Description automatically generated

The service of machine learning is still a service. The company most likely has a software monitoring process in place that can be employed. If the model runs in real-time, proper alerts and responsible persons on-call are required. We might experience data drift when the model receives data that it has not seen in training. Imagine users coming from a different age group, etcetera. Ultimately, the degradation of model quality is used to quantify drift. The simplest basic way to determine whether the model is working well is to compare the predictions to the actual values. The same metrics from the model training phase can be used, such as Precision/Recall for classification, RMSE for regression, and so on. If something goes wrong with the data quality or the real-world patterns, the metrics will fall.

A diagram of a model

Description automatically generated

Initial Model Fit Statistics

Understanding the statistics associated with the initial model fit is critical for measuring its success and laying the groundwork for future improvements when commencing predictive modeling. Several crucial statistics provide insight into how well your model matches the data and serve as a foundation for further development. Here are the metrics you've used to assess the efficacy of your model, as well as the insights they provide:

**The F1 Score:** It strikes a compromise between precision and recall. It's especially useful when working with skewed datasets, as it allows us to see how effectively our model classifies both positive and negative cases.

The F1 score of our best model, Random Forest is 0.9547.

**Area Under the ROC Curve:** The AUC-ROC measures the model's ability to differentiate between positive and negative classes at various probability levels. Higher AUC suggests greater discriminating power.

The AUC of the selected model i.e., Random Forest is 9.9935.

Risk Tiering

Risk tiering is a system for categorizing hazards based on their potential impact, likelihood of occurrence, and other pertinent characteristics. It is a way to more efficiently prioritize and manage risks by allocating them to different levels of severity. Based on the individual characteristics of each risk tier, airline companies can allocate resources, develop mitigation methods, and make educated decisions (Kiritz, 2019). Given that our optimal model is the Random Forest for this project, our focus will be on risk tiering concerning the benchmark performance of the model, specifically the AUC-ROC and F1 score. If the measurements deviate from their current values by 1% to 9%, steps will be taken to ensure the model's continuous operation. These activities will be determined by the risk tiering structure that has been defined below.

**No action:** No action will be needed if the drift in the benchmark is less than 3%. This will be categorized as low risk.

**Report:** This is the category of moderate risk, where it will be just reported if the drift is between the range of 3%-6%. The model will be tuned by using various optimization techniques.

**Refit:** If the drift is in the range of 6%-9% then the model will be refitted. This is categorized as high risk.

**Rebuild:** If the drift is greater than 9% then the model will be rebuilt as the model will be unable to make any prediction. This is categorized as an unacceptable risk.