**Business Analytics Capstone**

**Model Governance**

**Customer Churn Analysis for Jade Inc.**

**Submitted By:**

**Nimish Shaji Abraham**

**301312186**

**Advisor: Bilal Hassanzadah**

****Introduction****

The importance of model governance cannot be understated, especially for an organization like Jade Inc., which relies on predictive models for customer churn analysis. Model governance encompasses the framework, procedures, and policies that ensure the integrity, fairness, and reliability of the models throughout their lifecycle. In this report, we provide a detailed governance framework for Jade Inc., focusing on variable monitoring, model stability, risk tiering, and recommendations for ongoing model improvements.

The objective of this report is to outline a comprehensive model governance strategy for Jade Inc.'s churn model. This will include model validation procedures, variable-level monitoring, acceptable data ranges, model stability checks, risk tiering, and recommendations for future improvements. The governance framework aims to ensure that Jade’s churn prediction model performs optimally while maintaining compliance with data privacy regulations and ethical standards. Additionally, we explore risk mitigation strategies and monitoring practices essential for ensuring long-term model success.

****Validation: Monitoring and Governance****

The validation process for Jade Inc.'s churn prediction model followed industry best practices. We employed cross-validation techniques to ensure that the model performs well not only on training data but also on unseen data. Cross-validation splits the dataset into multiple folds and trains the model iteratively, testing on one fold while training on the others. This technique helps in identifying overfitting and ensures the model’s generalizability.

The key performance metrics assessed during the validation phase were accuracy, precision, recall, F1-score, and ROC-AUC score. Each of these metrics provides a unique insight into how well the model is able to distinguish between customers who are likely to churn and those who are not. Given the critical nature of churn prediction for Jade's customer retention strategies, the model’s accuracy across multiple validation sets provided confidence in its robustness and reliability.

In terms of governance, Jade Inc. adopted a framework that ensures model validation adheres to strict guidelines. Every model iteration is thoroughly documented, including the datasets used, validation metrics obtained, and any assumptions made during model building. This comprehensive documentation supports transparency and ensures accountability throughout the model's lifecycle.

**Variable Importance**

A set of important factors found through data exploration and feature engineering are the foundation of the predictive models created for Jade Inc. These variables include tenure which have a direct impact on churn, cashback amount, warehouse to home, and preferred login devices. To keep the model's prediction ability intact, these factors must be watched.

Variable-level monitoring is crucial to ensure that the model remains effective in making predictions. Any significant shifts in the distribution of these variables could indicate changes in customer behavior that need to be addressed. For example, a sudden increase in the frequency of complaints may signal that customers are becoming increasingly dissatisfied, and this would necessitate model recalibration to account for the new data trends.

A graph with a bar graph

Description automatically generatedBy monitoring key variables in real-time, Jade Inc. can proactively adjust its retention strategies and ensure that the churn model continues to capture the relevant predictors of customer churn accurately. The feature importance of all the variables for Random Forest is as follows:

**Distribution Consistency Checks**

To maintain model integrity, Jade Inc. employs regular consistency checks on the distribution of key variables. These checks ensure that the incoming data follows the same distribution as the data used to train the model. If there is a significant deviation, such as a shift in customer service interaction frequency or a rise in complaint rates, the model may require recalibration or even retraining.

Automated alerts are generated when the distribution of any key variable falls outside predefined thresholds. This ensures that Jade Inc. can take corrective action promptly, either by adjusting the model’s input preprocessing or retraining the model on updated data to maintain predictive accuracy.

Acceptable Ranges

**Defining Ranges**

For each of the key variables identified, acceptable ranges have been defined based on historical data analysis and industry benchmarks. These ranges provide a baseline for expected customer behavior and help in identifying any outliers or anomalies in the data. For example, customer tenure typically ranges between one month and 5 years for Jade’s customer base. Similarly, interaction frequency is expected to be within a certain range depending on the type of subscription the customer holds.

By establishing acceptable ranges for key variables, Jade Inc. ensures that the model remains focused on normal customer behavior while flagging any unusual patterns that may require further investigation. The acceptable ranges for the model is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Description | Data Type | Values |
| CustomerID | Unique customer ID | Numerical | * Numerical Values of the customer id |
| Churn | Whether a customer churned or not | Categorical | * Churn: Yes * Churn: No |
| Tenure | Tenure of customer in organization | Numerical | Number of Years a customer has been in the company for. Ranging from 6 months to 5 years. |
| PreferredLoginDevice | The preferred login device of the customer | Categorical | * MobilePhone * Computer * Phone: Relates to customers placing phone call orders |
| CityTier | Ranking given to a city. The lower the number the more modern the city, | Categorical | * 3: Low Tier city * 2: Medium Tier City * 1: High tier city |
| WareHouseToHome | Distance between the warehouse and the customers’ home | Numerical | Containing distance in kilometers all the way from 5 to 127. |
| PreferredPaymentMode | Preferred payment method of customer | Categorical | * Debit Card * Credit Card * Cash on Delivery * CC (Cash Credit) * COD (Credit on Delivery) * E-Wallets (Paypal) * UPI (Apple Pay) |
| Gender | Gender of customer | Categorical | * Male * Female |
| HourSpendOnApp | Total Number of Hours spent on the app | Numerical | * 1: Used the app for 1 hour * 2: Used the app for 2 hours * 3: Used the app for 3 hours |
| NumberOfDeviceRegistered | Total Number of Devices registed on Jade | Numerical | 1 to 6. Each number lists the number of devices registered. |
| PreferredOrderCategory | The preferred category which customers like to buy from | Categorical | * Fashion * Laptop and Mobile Accessory * Mobilephones * Grocery * Others |
| Satisfaction Score | Is the customer satisfied with us or not | Numerical | Satisfaction values from 1 to 5. 1 being the least satisfied and 5 being the most satisfied. |
| Marital Status | Provides information on the customers’ marital status | Categorical | * ‘No’: Customers have not churned * ‘Yes’: Customers have churned |
| NumberOfAddress | Provides information on the number of addresses a customer has | Numerical | Has values from 1 to all the way to 22. |
| Complain | Provides information on whether a complain has been raised by the customer or not | Categorical | * 0: No * 1: Yes |
| OrderAmountHikeFromlastYear | Provides information on the order amount hike from last year in percentage | Numerical | Consists data from 11% to 26% |
| CouponUsed | Total number of coupon used by the customer in the last month | Numerical | Consists numbers from 0 to 16 |
| OrderCount | Total number of orders being placed in the last month | Numerical | Consists data from 1 all the way to 16 |
| DaySinceLastOrder | Day Since last order by customer | Numerical | Contains values from 0 days to 46 |
| CashbackAmount | Average cashback amount received in the last month | Numerical | Contains values from 0 all the way to 324.99. |

All the variables that have to be used in the future must adhere to the above ranges.

**Monitoring for Anomalies**

Jade Inc. employs automated systems that continuously monitor whether key variables stay within their predefined acceptable ranges. When a variable deviates from its expected range, an alert is generated. These alerts prompt the data science and business teams to investigate the underlying cause of the anomaly and take necessary corrective action. This may involve updating the model’s input data, adjusting the preprocessing steps, or recalibrating the model to better handle outlier data points.

Monitoring for anomalies is crucial to maintaining the model's predictive power and ensuring that it continues to deliver reliable churn predictions in dynamic market conditions.

**Tolerance and Drift**

For Jade Inc.'s customer churn predictive models, establishing drift tolerance thresholds for key variables ensures the ongoing accuracy of the models. These thresholds define how much a variable can change before it significantly impacts the model’s predictive power. Monitoring these thresholds helps Jade stay responsive to shifts in customer behavior, ensuring the models remain effective over time. Below are the drift thresholds for key variables in Jade's customer churn analysis dashboard:

* **Tenure:** A shift of more than 10% in the distribution of customer tenure (from 6 months to 5 years) may indicate significant changes in customer loyalty or satisfaction patterns that require further investigation.
* **PreferredLoginDevice:** A change exceeding 15% in the distribution of preferred login devices (e.g., MobilePhone, Computer, Phone) could reflect shifts in how customers engage with Jade, potentially impacting retention strategies.
* **CityTier:** A change of more than 10% in the distribution of city tiers (Low, Medium, High) may signal broader changes in Jade's customer demographics and infrastructure preferences, impacting service and delivery strategies.
* **WareHouseToHome:** A shift of more than 10% in the average distance between the warehouse and customers' homes could suggest changes in delivery logistics or the geographical distribution of Jade’s customer base.
* **PreferredPaymentMode:** A fluctuation of more than 15% in the distribution of preferred payment modes (e.g., Debit Card, Credit Card, E-Wallets, UPI) could indicate shifting customer preferences in payment behavior, potentially requiring adjustments in payment options or promotions.
* **Gender:** A shift of more than 15% in the ratio of male to female customers could suggest significant demographic changes within Jade's customer base, warranting closer examination of engagement strategies.
* **HourSpendOnApp:** A shift of more than 10% in the distribution of hours spent on the app (from 1 to 3 hours) could indicate changes in customer engagement with the platform, potentially reflecting changes in the app's user interface or customer preferences.
* **NumberOfDeviceRegistered:** A change of more than 10% in the average number of devices registered per customer (ranging from 1 to 6) could signal shifts in how customers interact with Jade's services across multiple devices.
* **PreferredOrderCategory:** A change exceeding 10% in the distribution of preferred order categories (e.g., Fashion, Mobile Phones, Groceries) may indicate changing customer preferences and demand, affecting inventory and marketing strategies.
* **Satisfaction Score:** A shift of more than 10% in the average customer satisfaction score (from 1 to 5) could suggest significant changes in customer sentiment and experience with Jade’s services, potentially requiring adjustments to service offerings or customer support.
* **Marital Status:** A change of more than 15% in the ratio of married to single customers could indicate demographic shifts that may affect customer behavior and purchasing patterns.
* **NumberOfAddress:** A shift of more than 10% in the average number of addresses per customer (from 1 to 22) could suggest changes in customer mobility or residency patterns that may impact delivery logistics.
* **Complain:** A fluctuation of more than 10% in the proportion of customers raising complaints (Yes/No) could signal shifts in customer satisfaction or service quality, necessitating closer attention to customer support.
* **OrderAmountHikeFromlastYear:** A deviation of more than 10% in the average percentage increase in order amount from the previous year (11% to 26%) could indicate changes in customer spending behavior or pricing strategies.
* **CouponUsed:** A change exceeding 10% in the average number of coupons used per customer (ranging from 0 to 16) may reflect shifts in customer engagement with promotions and discounts, potentially impacting marketing effectiveness.
* **OrderCount:** A shift of more than 10% in the distribution of the number of orders placed by customers in the last month (from 1 to 16) could indicate changes in purchasing patterns or customer demand.
* **DaySinceLastOrder:** A fluctuation of more than 10% in the distribution of days since the last order (from 0 to 46 days) could suggest changes in customer retention or purchasing frequency.
* **CashbackAmount:** A shift of more than 10% in the average cashback amount received by customers (from $0 to $324.99) could signal changes in customer spending or engagement with cashback incentives, potentially affecting customer retention strategies.

By establishing and monitoring these drift tolerance thresholds, Jade Inc. can maintain the effectiveness of its predictive models, ensuring that they continue to provide accurate insights into customer churn even as customer behaviors and market trends evolve.

Model Health and Stability

**Key Performance Metrics**

To ensure that Jade Inc.’s churn prediction model remains effective over time, regular assessments of its health and stability are conducted. Key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score are tracked continuously to gauge the model’s performance. If any of these metrics show a significant decline, the model may need to be retrained or recalibrated.

For example, if the accuracy drops below 85%, or if the recall for churned customers falls significantly, this could indicate that the model is no longer capturing the appropriate predictors of churn. In such cases, the model is re-evaluated and adjusted as necessary.

**Recalibration and Stability Monitoring**

Recalibration involves adjusting the model’s parameters and retraining it with updated data to ensure that it stays aligned with current customer behavior. Stability monitoring ensures that the model continues to perform well across different time periods and datasets. This helps prevent issues such as model drift, where the model’s performance degrades over time as the underlying data changes.

Regular recalibration and stability checks are a cornerstone of Jade Inc.’s model governance strategy, ensuring that the churn model remains a reliable tool for predicting customer churn and informing business decisions.

Risk Tiering

**Assessing Model Risk**

Given the critical role that customer retention plays in Jade Inc.’s business strategy, the churn prediction model is classified as a high-risk model. High-risk models typically have significant business impact, and any failure or underperformance can result in substantial financial losses. As a result, high-risk models require more rigorous validation, monitoring, and governance.

To assess the risk associated with the churn model, we evaluated several factors, including the complexity of the model, its potential business impact, and the risks associated with incorrect predictions. The assessment indicated that any significant drop in model performance could lead to a higher customer churn rate, thereby affecting the company’s revenue and reputation.

**Mitigation Strategies**

To mitigate the risks associated with the churn model, several strategies have been implemented. These include regular retraining schedules, fallback models, and contingency plans. Regular retraining ensures that the model stays current with changing customer behavior, while fallback models serve as a backup in case the primary model underperforms.

Additionally, Jade Inc. has established a manual review process for high-risk predictions. In cases where the model predicts a high likelihood of churn, these predictions are flagged for manual review by a customer retention specialist. This ensures that Jade can take proactive steps to prevent churn, even in cases where the model’s predictions may be uncertain.

Initial Model Fit Statistics

When evaluating the best predictive model developed for Jade Inc. out of logistic regression (full, forward, backward, stepwise), decision tree, and random forest, the key evaluation metrics of ROC-AUC, accuracy, and F1-Score have been employed to assess each model's fit and performance. After comparing the results, the random forest model emerged as the optimal predictive model due to its strong performance across all metrics.

From the perspective of the ROC-AUC score, the random forest model achieved a value of 0.91, indicating that the model is 91% effective in distinguishing between customers who are likely to churn and those who are not likely to churn within Jade's customer base. Additionally, the random forest model attained an accuracy score of 97%, meaning that 91% of the model's predictions (both churn and non-churn) were correct. Lastly, the random forest model produced an F1-Score of 0.98 for No churn and 0.91 for Churn, demonstrating that the model is effective in correctly identifying customers that are likely to churn, balancing precision and recall effectively in the context of customer churn prediction for Jade Inc.

Recommendations for Future Studies

**Sample Size:**

* **Importance:** Adequate sample size is essential for the statistical robustness of the models. Smaller sample sizes could lead to overfitting, reduced generalization, and model bias. Expanding the sample size allows for more granular customer segmentation and improves the predictive power across diverse customer groups.
* **Action:** Future research should prioritize data collection efforts to increase the sample size, particularly in underrepresented segments of the customer base.

**Non-Linearity:**

* **Importance:** Many of Jade's customer behaviors and interactions exhibit non-linear relationships that traditional models may not fully capture. Recognizing and modeling these non-linearities will improve the prediction of customer churn.
* **Action:** Research should focus on exploring non-linear relationships between variables, particularly by employing machine learning models such as Random Forest and Neural Networks, which can capture these complexities.

**Modeling Techniques:**

* **Importance:** While traditional modeling techniques have proven effective, incorporating more advanced techniques such as ensemble learning, neural networks, and gradient boosting methods could further enhance predictive performance.
* **Action:** Future model development should experiment with diverse modeling techniques to identify those best suited for Jade's specific customer churn scenarios, optimizing accuracy and interpretability.

**Missing Values:**

* **Importance:** Missing values can reduce the accuracy of models and lead to biased results. Handling missing data through imputation or other techniques ensures the model remains reliable even when full data sets are unavailable.
* **Action:** Further research should investigate robust methods of dealing with missing values, particularly using domain-specific imputation methods or predictive modeling techniques designed to handle incomplete datasets.

**No Outlier Reduction:**

* **Importance:** Outliers often provide valuable insights into extreme customer behavior and should not be reduced or excluded from the dataset unless shown to be erroneous. By focusing on outlier detection and handling, Jade can better understand atypical customer patterns and their impact on churn.
* **Action:** Future research should explore methods that accommodate outliers within the data rather than removing them, ensuring that these data points are leveraged to improve model predictions.

References

Awan, A. (2023, September). *The Curse of Dimensionality in Machine Learning: Challenges, Impacts, and Solutions*. Datacamp; Datacamp. https://www.datacamp.com/blog/curse-of-dimensionality-machine-learning

DHL. (2024, January 25). *Customer Satisfaction In E-Commerce - DHL Express LK*. Dhl.com. https://www.dhl.com/discover/en-lk/e-commerce-advice/e-commerce-best-practice/customer-satisfaction-in-e-commerce

Lin, G. T. R., & Sun, C. (2009). Factors influencing satisfaction and loyalty in online shopping: an integrated model. *Online Information Review*, *33*(3), 458–475. https://doi.org/10.1108/14684520910969907

Matuszelański, K., & Kopczewska, K. (2022). Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(1), 165–198. https://doi.org/10.3390/jtaer17010009

Rudälv, A. (2022). *Predicting Customer Churn in E-Commerce Using Statistical Modeling and Feature Importance Analysis A Comparison of Random Forest and Logistic Regression Approaches*.