301335423

# Vortex Telecommunications Data Governance

Vinay Beesa Gnaneshwar

8-15-2024

# Contents

# 1. Data Governance Strategy

Objective: The main aim of Vortex's Data Governance Strategy is to guarantee optimal data integrity, security, and compliance, thereby improving the efficacy of predictive models designed to minimize customer churn. This strategy enables Vortex to utilize data-driven insights to retain customers identified as high-risk through sophisticated machine learning techniques, especially the Random Forest model.

Scope: The governance strategy covers the entire data lifecycle at Vortex. This includes gathering data from operational systems, preparing it, training and evaluating models, and using the insights gained to improve customer retention strategies.

Stakeholders: The main stakeholders involved are the Data Governance Officer, Data Stewards, Data Scientists and Business Analysts who are focused on predicting customer churn as well as IT Security, Customer Service Managers and the Executive Leadership team at Vortex.

# 2. Policies & Principles

## Data Quality:

Accuracy & Completeness:

- Emphasize the importance of ensuring high precision and thoroughness in essential data elements like Contract Type, Satisfaction Score, Monthly Charges, and Tenure in Months and Churn Reason which Random Forest model has recognized these factors as key indicators of churn.
- Consistently review the Contract Type field which is recognized by the model a key feature to ensure that the information remains current and accurately represents the customer's contract status.

Consistency:

- Implement strict data formatting standards across all datasets, including CustomerChurn, Telco_customer_churn_demographics, and Telco_customer_churn_status, to ensure that key variables like Churn Score, Internet Type, and Satisfaction Score are consistently represented and interpreted across different models and reports.

### Data Security & Privacy:

Encryption:

- Encrypt all sensitive customer data, especially fields like Payment Method, Total Revenue and Total Charges.

Access Control:

- Restrict access to sensitive datasets containing fields like Customer Status and Churn Reason, ensuring that only the Data Science team and designated stakeholders have access to this information.

Data Retention:

- Retain historical data, particularly past values of Satisfaction Score, Churn Score and Monthly Charges. This allows for continuous improvement of the predictive models by providing a richer dataset for training and validation.

## 3. Organization

### Roles & Responsibilities:

Data Steward:

- Oversees the quality of key variables identified in your analysis, such as Contract Type, Number of Referrals, and Internet Type. They ensure that these data points are accurately maintained and reflect the current state of customer interactions.

Data / Business Analysts:

- Focus on refining the Random Forest model, identified as the most effective for predicting customer churn. They ensure that the model remains accurate and up-to-date as new data becomes available, as highlighted in your project's model comparison section.

IT & Security Teams:

- Manage the secure storage and access of critical customer data.

Customer Service Managers:

- Use insights from the Random Forest model to guide targeted retention efforts, focusing on customers with high churn probabilities identified.

# 4. Processes

## Data Collection:

Data Sources:

- Regularly collect and integrate data from Vortex's operational systems, including customer interaction logs, billing systems, and service usage records. For example, ensure that the Monthly Charges and Total Revenue fields are accurately recorded and reflect the customer's latest transactions, as demonstrated in the summary of descriptive statistics.

Data Integration:

- Data from below datasets are merged. By integrating these datasets, Vortex can create a comprehensive view of each customer, enabling more accurate predictions of churn. The merged data is then pre-processed for model training, ensuring that all variables are properly formatted and free from discrepancies.

  *CustomerChurn*
  *Telco_customer_churn*
  *Telco_customer_churn_demographics*
  *Telco_customer_churn_location*
  *Telco_customer_churn_status*

### Data Cleaning & Pre-processing:

Dropping of Variables:

- View documentation to view all list of variables to be dropped. But dropping of variables in key to maintaining data cleanliness.

Missing Values:

- Implement targeted imputation strategies for missing values.
- Missing values in Churn_Reason were imputed with "Customer Active."
- Missing values in Offer were filled with "Unknown Offer."
- Missing values in Internet_Type were imputed with "No Internet Service."
- Missing values in Churn_Category were filled with "Customer Active."

Outlier Detection:

- Outliers were only viewed and not imputed for to retain originality of the dataset.

Version Control:

- Maintain version control for the Random Forest model, documenting changes in feature selection, hyper parameters, and pre-processing steps.

## 5. Tools & Technology

### Data Storage:

Secure Cloud Storage:

- Store critical datasets like CustomerChurn, Telco_customer_churn_status and Telco_customer_churn_demographics in a secure cloud-based storage solution that supports encryption, automated backups, and role-based access controls.

### Modelling & Analysis Tools:

Python Libraries: Use Python libraries listed below

- Pandas (import pandas as pd)
- NumPy (import numpy as np)
- Matplotlib (import matplotlib.pyplot as plt)
- Seaborn (import seaborn as sns)
- Re (import re)
- Tabulate (from tabulate import tabulate)
- Folium (import folium)
- Folium Plugins (from folium.plugins import HeatMap)
- SciPy (from scipy import stats)
- SciPy Stats (from scipy.stats import norm)
- Statsmodels (import statsmodels.api as sm)
- Scikit-learn
  - from sklearn.linear_model import LogisticRegression
  - from sklearn.neighbors import KNeighborsClassifier
  - from sklearn.ensemble import RandomForestClassifier
  - from sklearn.neural_network import MLPClassifier
  - from sklearn.model_selection import train_test_split
  - from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay, classification_report, roc_auc_score
  - from sklearn.metrics import precision_recall_curve, auc
  - from sklearn.metrics import roc_curve
  - from sklearn.calibration import calibration_curve
  - from sklearn.inspection import PartialDependenceDisplay
  - from sklearn.inspection import permutation_importance

## 6. Governance Controls

### Key Performance Indicators (KPIs):

- Accuracy
- Precision
- Recall
- F1-Score
- ROC AUC Score

Model Performance Metrics:

| Metric | Current Value | Acceptable Minimum |
|---|---|---|
| Accuracy | 97.60% | 95% |
| Precision | 100% (Churn), 95% (Non-Churn) | 90% |
| Recall | 94% (Churn), 100% (Non-Churn) | 85% |
| F1-Score | 97% (Churn), 99% (Non-Churn) | 85% |
| ROC AUC Score | 0.97 | 0.90 |

## 7. Implementation Model

### Phased Implementation:

• Phase 1: Establish and enforce data governance policies, focusing on the critical features identified in analysis.

• Phase 2: Implement data quality controls and secure storage solutions ensuring that customer data is protected and accurately maintained.

• Phase 3: Deploy the Random Forest model integrating it into Vortex's decision-making processes.

## 8. Continuous Governance Improvement

### Feedback Loop:

• Regular Reviews: Schedule quarterly reviews of data governance practices, focusing on areas like data quality and model performance. If the model's recall or precision drops, investigate whether data quality issues or changes in customer behaviour are the root cause.

• Model Retraining: Retrain the Random Forest model as new customer data becomes available. This continuous improvement process ensures the model adapts to changes in customer behaviour and remains effective in predicting churn.

## Training & Awareness:

- Employee Training: Conduct regular training sessions for Vortex staff on data governance best practices, emphasizing the importance of maintaining high data quality.

- Governance Updates: Keep all stakeholders informed about updates to governance practices, particularly if changes are made to the data management processes or model parameters.