Assessment cover sheet

Your name: Jeevan Kokkadan Johny

Specific feedback request

If you would like specific feedback on any part of your submission, please note it below (max 75 words):

Example: 'I wasn't sure how best to represent the data on Slide 6. I used a bar chart but struggled with labelling the axes clearly.' or 'I had trouble with line 34 of my code, any tips for improvement?'

Mentem by UNSW

Instructions:

1. Include this cover sheet as the first page of your submission.

2. Convert your document(s) to PDF before uploading.

3. Submit this PDF along with any other deliverables specified in the assessment brief.

Modelling Customer Churn

Report

Jeevan Kokkadan Johny

Customer Churn Prediction

Comparison of Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting Models

Model	ROC/AUC	Precision	Recall	F1 Score	Training Accuracy	Testing Accuracy
Logistic Regression	0.778	0.552	0.201	0.295	0.811	0.811
Decision Tree	0.685	0.471	0.511	0.490	1.000	0.791
Random Forest	0.854	0.769	0.473	0.586	1.000	0.869
Gradient Boosting	0.870	0.741	0.481	0.583	0.874	0.865

Summary:

•**Gradient Boosting** has the highest ROC/AUC score (0.870), indicating the best overall performance in distinguishing between classes.

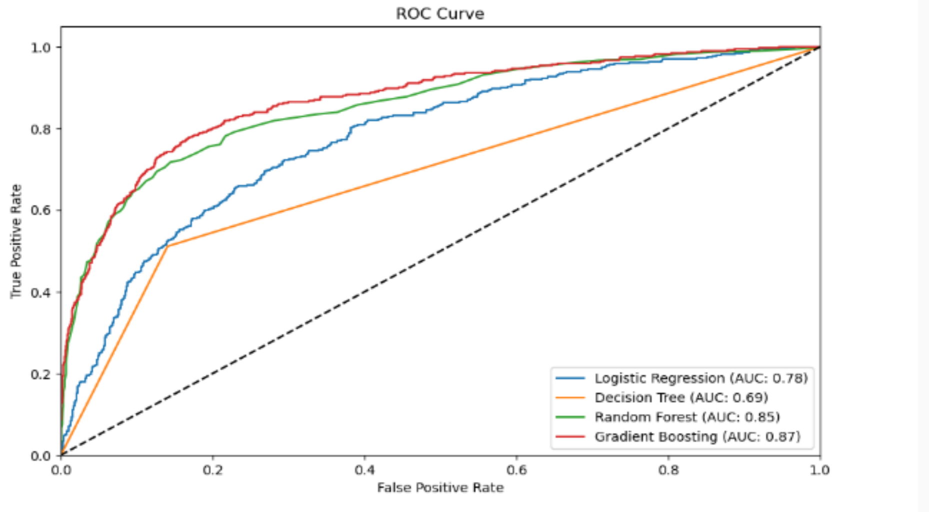
•**Random Forest** also performs well with a high ROC/AUC score (0.854) and the highest Precision (0.769).

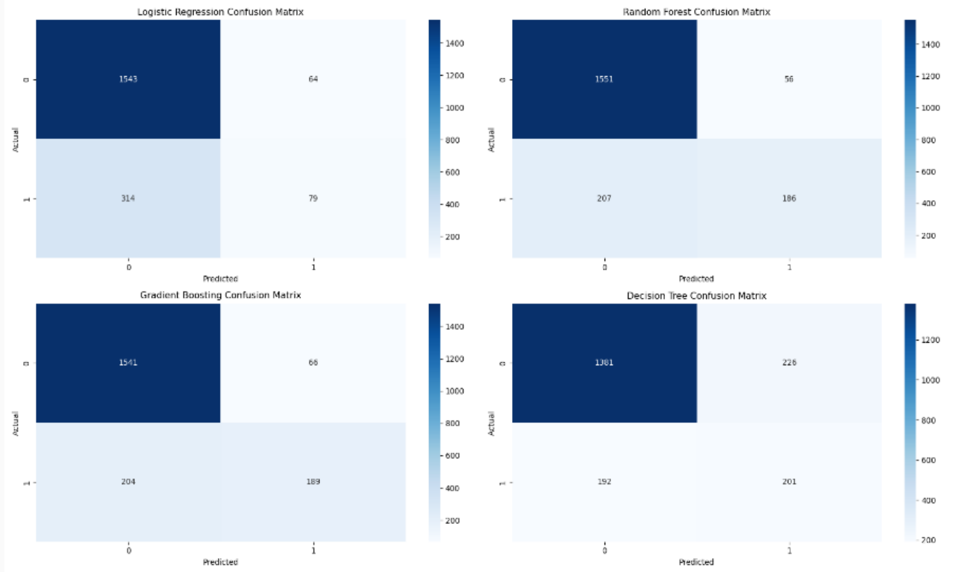
•**Logistic Regression** has moderate performance but lower Recall (0.201), indicating it misses more true positives.

•**Decision Tree** shows signs of overfitting with perfect Training Accuracy (1.000) but lower Testing Accuracy (0.791).

Evaluating Model Performance using ROC/AUC, Precision and Recall

- The ROC curve visualization shows that Gradient Boosting and Random Forest models outperform Logistic Regression and Decision Tree models in distinguishing between customers who will churn and those who will not. Gradient Boosting has the highest overall performance, followed closely by Random Forest, indicating their superior predictive capabilities.
- The confusion matrices for Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting models show their performance in predicting customer churn. Logistic Regression has moderate precision, meaning it correctly identifies 55.2% of the positive predictions. However, its recall is low at 20.1%, indicating it misses a significant number of true positives. The Decision Tree model has lower precision (47.1%) compared to Logistic Regression but higher recall (51.1%). This means it identifies more true positives but also has more false positives. Random Forest has the highest precision (76.9%) among the models, indicating it makes fewer false positive predictions. Its recall is moderate at 47.3%, meaning it identifies a fair number of true positives. Gradient Boosting also has high precision (74.1%) and moderate recall (48.1%). It performs well in correctly identifying positive predictions while maintaining a balance between precision and recall.



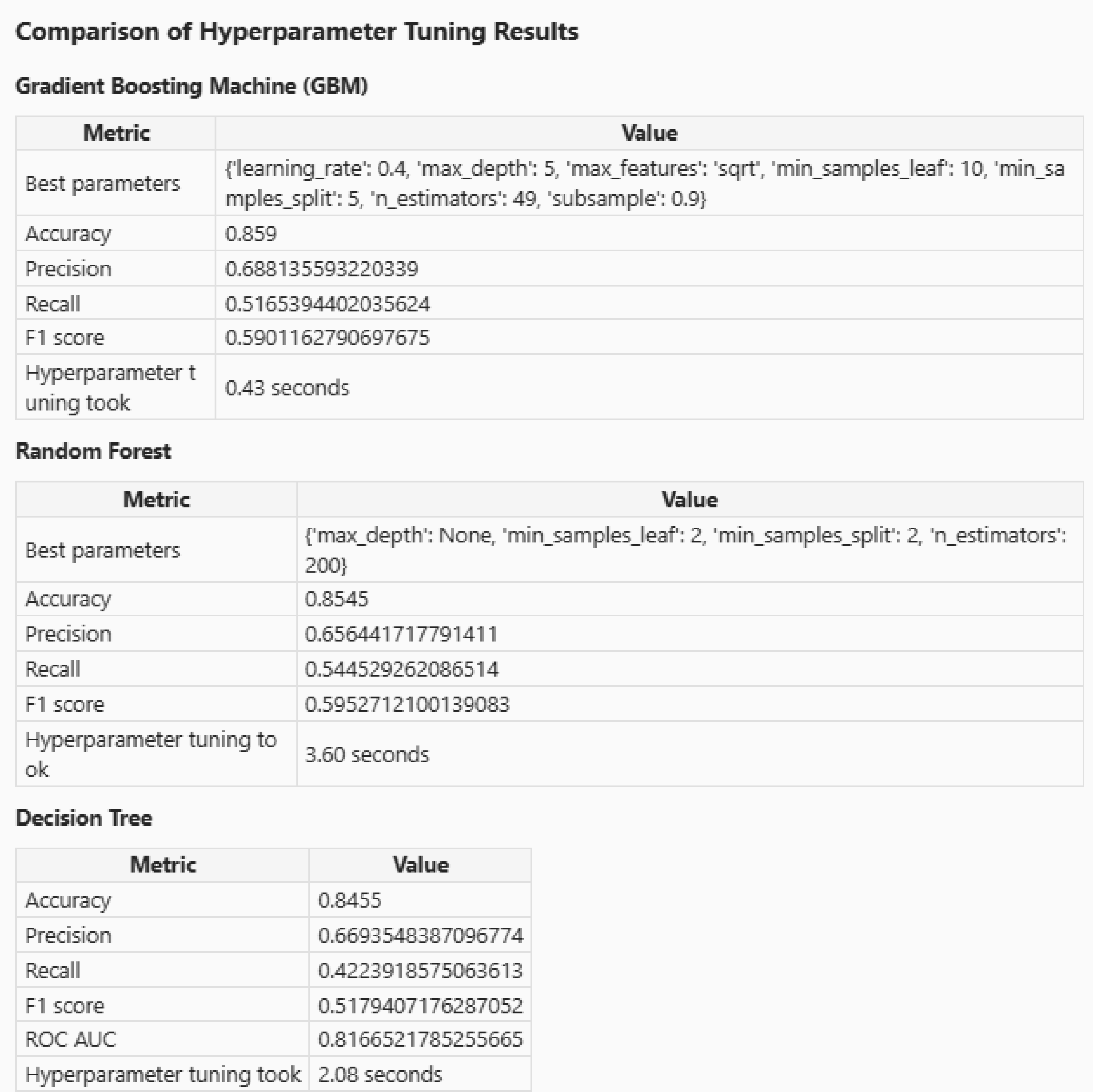


Hyperparameter Tuning for Different Models

Gradient Boosting Machine (GBM) has the highest accuracy (0.859) and precision (0.688), making it a strong candidate for use. It also has a relatively short hyperparameter tuning time (0.43 seconds).

Random Forest has a slightly lower accuracy (0.8545) and precision (0.656), but it has a higher recall (0.545) and F1 score (0.595) compared to GBM. However, it takes longer for hyperparameter tuning (3.60 seconds).

Decision Tree has the lowest accuracy (0.8455) and recall (0.422), but it has a higher precision (0.669) compared to Random Forest. It also has a relatively short hyperparameter tuning time (2.08 seconds).



Comparison of Hyperparameter Tuning Results

Gradient Boosting Machine (GBM)

Metric	Value
Best parameters	{'learning_rate': 0.4, 'max_depth': 5, 'max_features': 'sqrt', 'min_samples_leaf': 10, 'min_samples_split': 5, 'n_estimators': 49, 'subsample': 0.9}
Accuracy	0.859
Precision	0.688135593220339
Recall	0.5165394402035624
F1 score	0.5901162790697675
Hyperparameter tuning took	0.43 seconds

Random Forest

Metric	Value
Best parameters	{'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}
Accuracy	0.8545
Precision	0.656441717791411
Recall	0.544529262086514
F1 score	0.5952712100139083
Hyperparameter tuning took	3.60 seconds

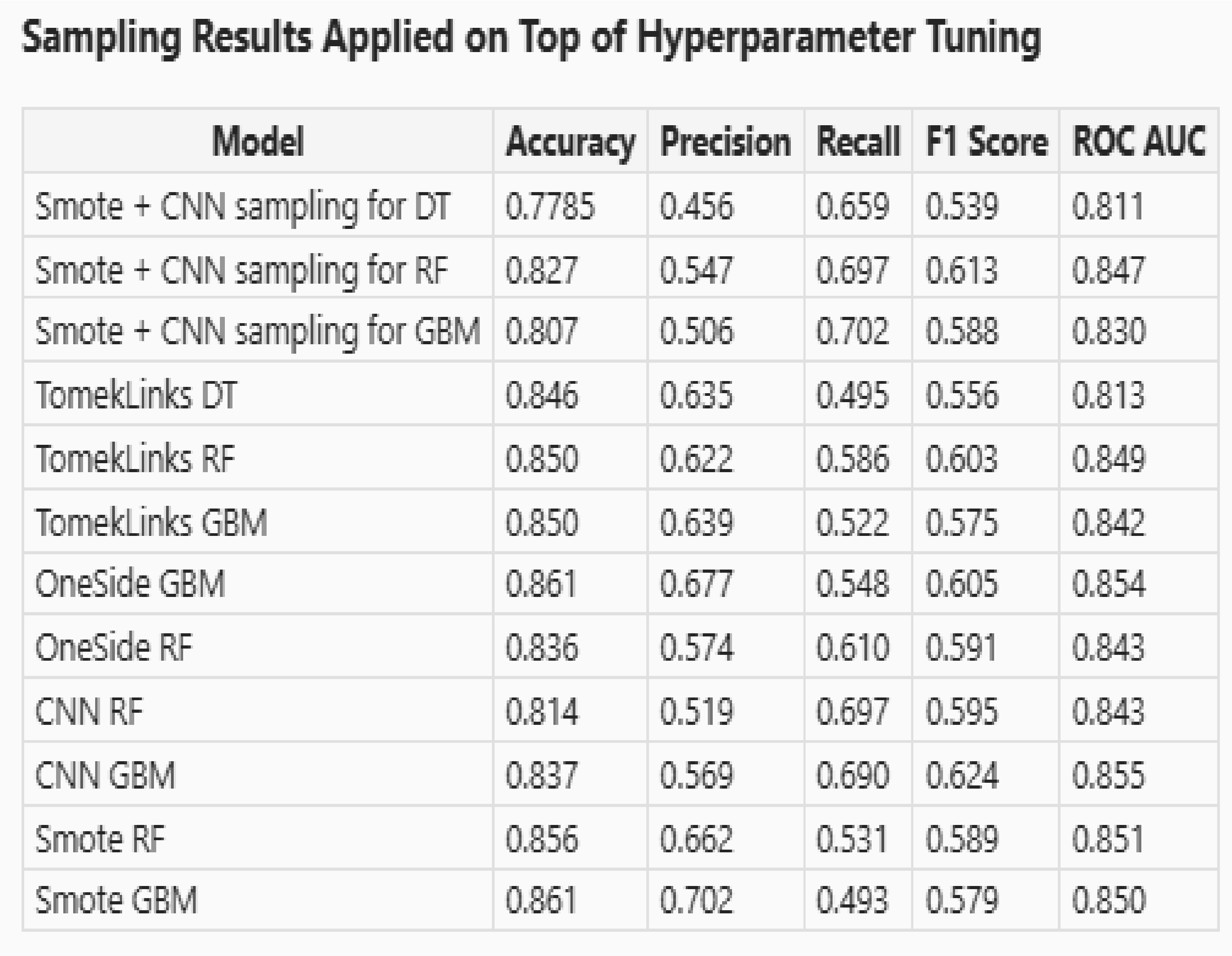
Decision Tree

Metric	Value
Accuracy	0.8455
Precision	0.6693548387096774
Recall	0.4223918575063613
F1 score	0.5179407176287052
ROC AUC	0.8166521785255665
Hyperparameter tuning took	2.08 seconds

Data Analysis and Feature Engineering Details

Data Leakage	The correlation between the Complain and Exited variables in the dataset is very high. This strong correlation indicates that customers who have filed complaints are highly likely to exit. To prevent data leakage, the Complain variable was excluded from model training.
Dropping Unnecessary Columns	Dropped columns that are not useful for prediction: Row number, Customerid, Surname.
Binning	Binned the following variables: Age, NumOfProducts, Satisfaction Score.
Feature Engineering by creating separate bins	Created separate bins for: Age (age_Group_age group_0-20, age_Group_age group_21-30, age_Group_age group_31-40, age_Group_age group_41-50, age_Group_age group_51-60, age_Group_age group_61-70, age_Group_age group_70+), NumOfProducts (numofproducts_Binned_1 product, numofproducts_Binned_2 products, numofproducts_Binned_3 products, numofproducts_Binned_4+ products), Satisfaction Score (satisfaction_score_Binned_satisfaction_1, satisfaction_score_Binned_satisfaction_2, satisfaction_score_Binned_satisfaction_3, satisfaction_score_Binned_satisfaction_4, satisfaction_score_Binned_satisfaction_5). Excluded original Age, NumOfProducts, and Satisfaction Score from model training.
Encoding	Applied one-hot encoding to all categorical columns. For Gender, applied dummy encoding. For other categorical fields, applied one-hot encoding.
Scaling	Applied standard scaling to the following fields to standardize the values: Balance, Estimated Salary, Credit Score.

Data Sampling Techniques



Sampling Results Applied on Top of Hyperparameter Tuning

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Smote + CNN sampling for DT	0.7785	0.456	0.659	0.539	0.811
Smote + CNN sampling for RF	0.827	0.547	0.697	0.613	0.847
Smote + CNN sampling for GBM	0.807	0.506	0.702	0.588	0.830
TomekLinks DT	0.846	0.635	0.495	0.556	0.813
TomekLinks RF	0.850	0.622	0.586	0.603	0.849
TomekLinks GBM	0.850	0.639	0.522	0.575	0.842
OneSide GBM	0.861	0.677	0.548	0.605	0.854
OneSide RF	0.836	0.574	0.610	0.591	0.843
CNN RF	0.814	0.519	0.697	0.595	0.843
CNN GBM	0.837	0.569	0.690	0.624	0.855
Smote RF	0.856	0.662	0.531	0.589	0.851
Smote GBM	0.861	0.702	0.493	0.579	0.850

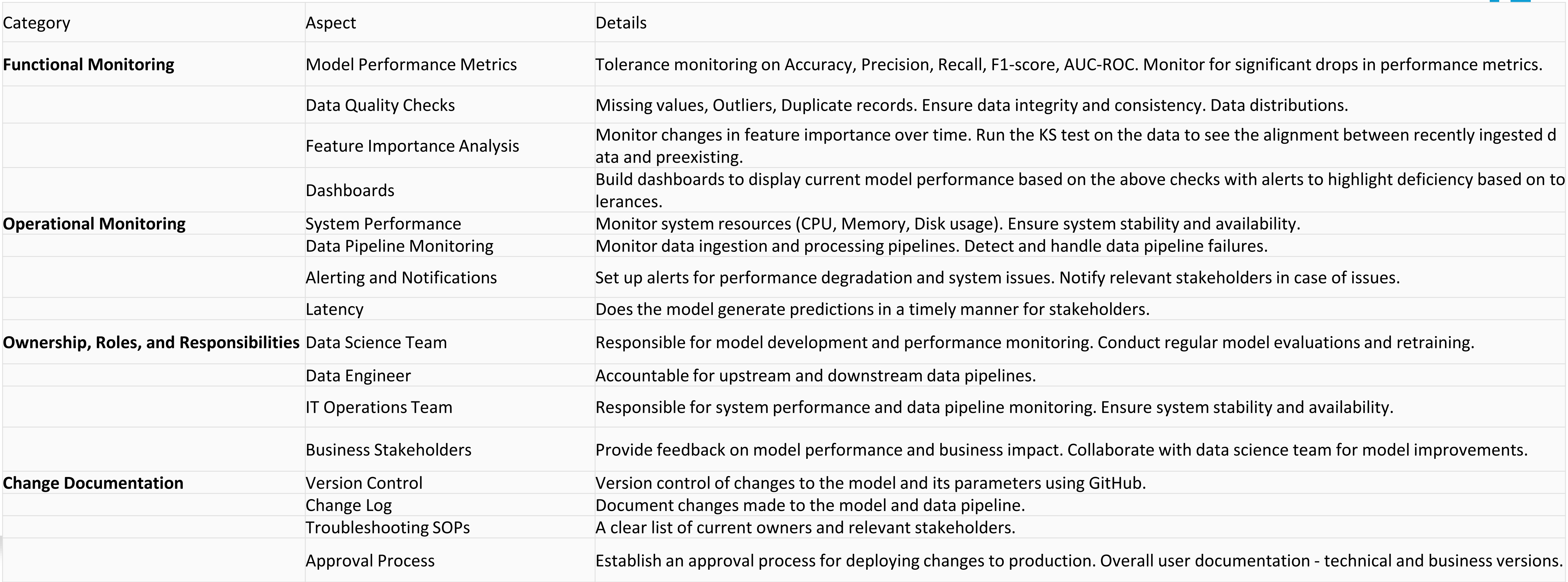
OSS GBM has the highest accuracy (0.861) and a good balance of precision (0.677), recall (0.548), and F1 score (0.605), making it a strong candidate for use.

Smote GBM also has high accuracy (0.861) and precision (0.702), but lower recall (0.493) and F1 score (0.579) compared to OSS GBM.

Tomek Links RF and **Smote RF** have good performance metrics, but OSS GBM outperforms them in terms of accuracy and precision

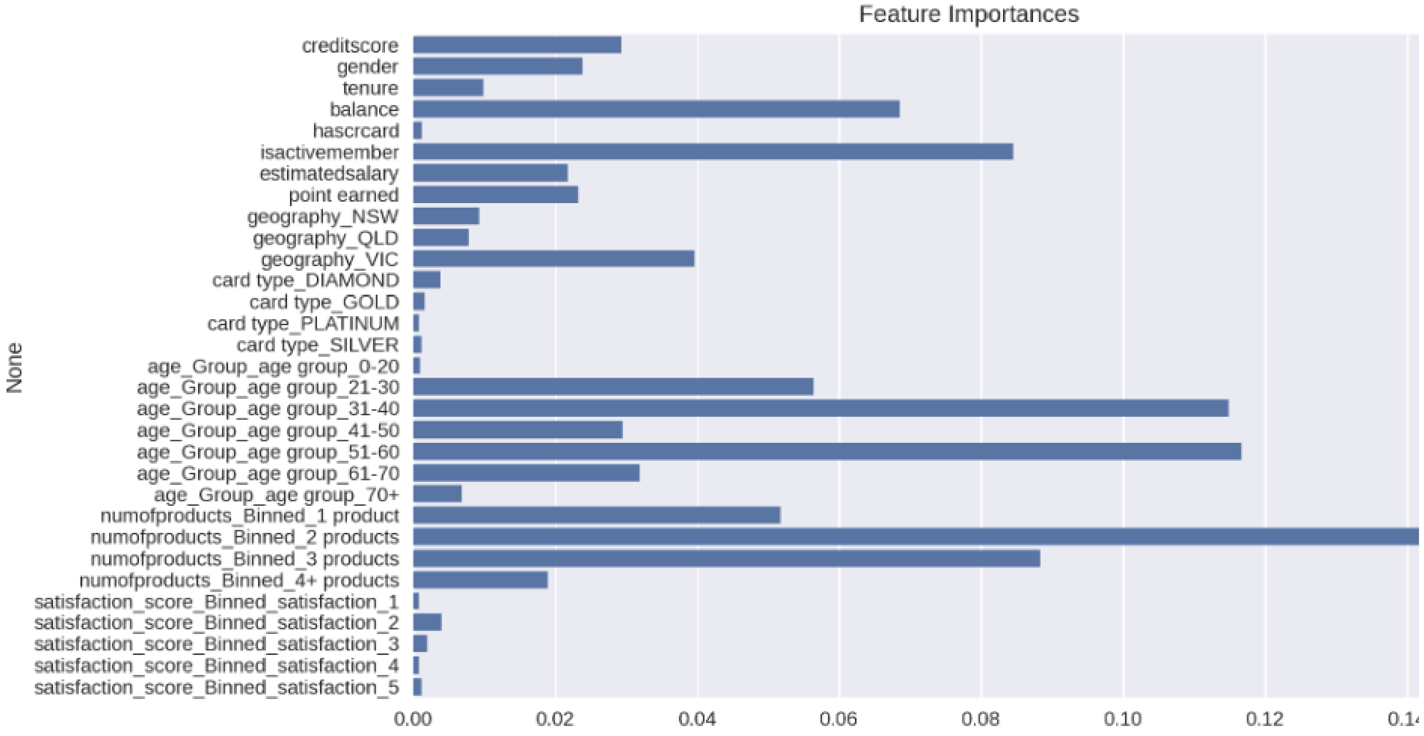
OSS GBM has the highest accuracy (0.861) and a good balance of precision (0.677), recall (0.548), and F1 score (0.605), making it a strong candidate for use.
CNN GBM also performs well with high accuracy (0.8365) and F1 score (0.624), but OSS GBM outperforms it in terms of accuracy and precision.

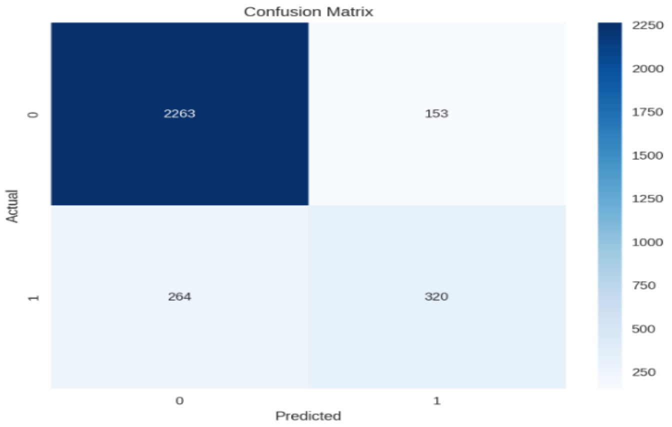
Monitoring Plan for Churn Prediction Model

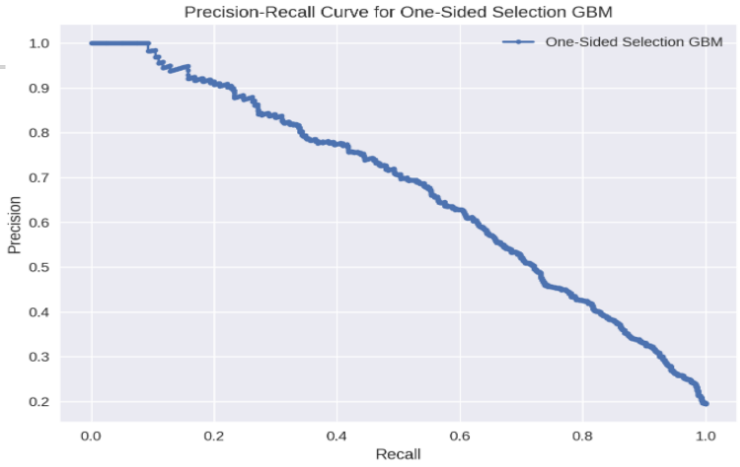


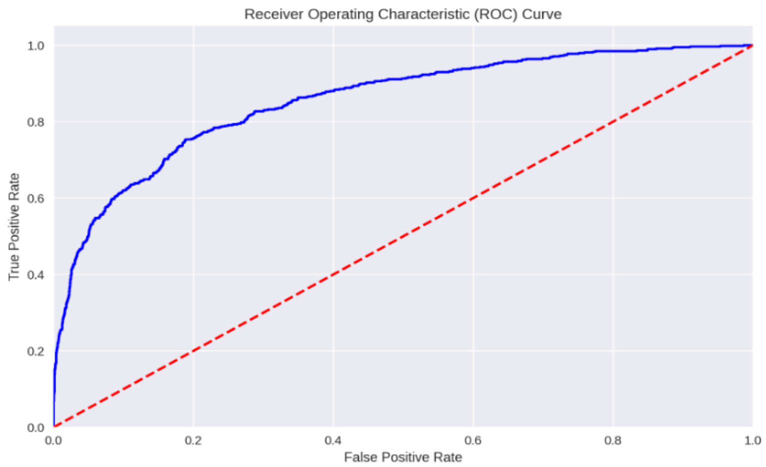
Category	Aspect	Details
Functional Monitoring	Model Performance Metrics	Tolerance monitoring on Accuracy, Precision, Recall, F1-score, AUC-ROC. Monitor for significant drops in performance metrics.
	Data Quality Checks	Missing values, Outliers, Duplicate records. Ensure data integrity and consistency. Data distributions.
	Feature Importance Analysis	Monitor changes in feature importance over time. Run the KS test on the data to see the alignment between recently ingested data and preexisting.
	Dashboards	Build dashboards to display current model performance based on the above checks with alerts to highlight deficiency based on tolerances.
Operational Monitoring	System Performance	Monitor system resources (CPU, Memory, Disk usage). Ensure system stability and availability.
	Data Pipeline Monitoring	Monitor data ingestion and processing pipelines. Detect and handle data pipeline failures.
	Alerting and Notifications	Set up alerts for performance degradation and system issues. Notify relevant stakeholders in case of issues.
	Latency	Does the model generate predictions in a timely manner for stakeholders.
Ownership, Roles, and Responsibilities	Data Science Team	Responsible for model development and performance monitoring. Conduct regular model evaluations and retraining.
	Data Engineer	Accountable for upstream and downstream data pipelines.
	IT Operations Team	Responsible for system performance and data pipeline monitoring. Ensure system stability and availability.
	Business Stakeholders	Provide feedback on model performance and business impact. Collaborate with data science team for model improvements.
Change Documentation	Version Control	Version control of changes to the model and its parameters using GitHub.
	Change Log	Document changes made to the model and data pipeline.
	Troubleshooting SOPs	A clear list of current owners and relevant stakeholders.
	Approval Process	Establish an approval process for deploying changes to production. Overall user documentation - technical and business versions.

Customer Churn Model Visualizations









Ethical, Bias, and Privacy Considerations

Ethical Implications:

Impact on Statistical Minority Groups:

Ensuring Diverse Representation: Ensure diverse representation in the training data by randomizing and equally distributing data among different customer groups.

Fairness: Ensure the model does not disproportionately affect minority groups.

Transparency: Clearly communicate how the decision-making process works.

Accountability: Establish mechanisms for addressing grievances and correcting errors.

Regular Audits: Conduct periodic reviews to identify and mitigate any adverse impacts on minority groups.

Governance Framework: Establish a governance framework to oversee ethical issues and address ethical concerns promptly.

Bias Risks:

Sampling Bias: Ensure that the data collected is representative of the entire customer base, including minority groups. Avoid over-representation or under-representation of any group.

Historical Bias: Be aware of any historical biases present in the data that could influence the model's predictions. This includes biases in credit scores, geographical location, and other demographic factors.

Feature Selection: Carefully select features that do not introduce bias. For example, avoid using features that are proxies for protected attributes (e.g. gender).

Model Evaluation: Regularly test the model for bias and adjust as necessary to avoid overfitting or underfitting.

Compliance:

Regulatory Compliance: Ensure that the model complies with relevant data protection regulations such as GDPR, CCPA, and other local privacy laws. Regularly review and update compliance measures to stay current with legal requirements.

Data Privacy:

Data Anonymization: Anonymize personal data to protect customer identities. This includes removing or encrypting personally identifiable information (PII) such as names, addresses, and customer IDs.

Data Minimization: Collect only the data necessary for the model's purpose. Avoid collecting excessive or irrelevant data that could increase privacy risks.

Data Security:

Access Controls: Implement strict access controls to ensure that only authorized personnel can access sensitive data. Use role-based access control (RBAC) to limit access based on job responsibilities.

Data Encryption: Encrypt data both at rest and in transit to protect it from unauthorized access and breaches.

Conclusion and Proposed Solution

Problem Statement

- A leading bank has been facing an increase in customer churn, significantly affecting their profitability and business. To address this issue, a comprehensive model monitoring plan has been developed.

Approach

- Compared different sampling techniques and hyperparameter tuning methods to identify the best model for predicting customer churn. The following sampling techniques were used:

1. **SMOTE + CNN**
2. **Tomek Links**
3. **One-Sided Selection**
4. **CNN**
5. **SMOTE**

- The models evaluated were:

1. **Decision Tree (DT)**
2. **Random Forest (RF)**
3. **Gradient Boosting Machine (GBM)**

Proposed Solution

- Based on the comparison, the **One-Sided Selection GBM** model is recommended as it has the highest accuracy (86.1%) and provides a good balance between precision, recall, and F1 score:
- **Precision**: 67.65%
- **Recall**: 54.79%
- **F1 Score**: 60.55%
- **ROC AUC**: 85.44%
- While the CNN model has a slightly higher ROC AUC, the One-Sided Selection GBM model provides better overall performance in terms of accuracy and precision.