Capstone Project - PGP-DSBA :- Customer Churn

1. Introduction

Customer churn is a critical challenge for the DTH provider due to intense market competition. Losing accounts, which represent multiple customers, significantly impacts the customer base and revenue. To proactively address this, the company requires a predictive model to identify at-risk accounts. This will enable targeted, cost-effective interventions, shifting from a reactive to a data-driven approach. The goal is to improve customer retention and protect revenue by implementing a churn prediction model and strategically segmented campaigns.

Shape & Info of Data Set

Shape of the data

```
[ ] df.shape
```

```
(11260, 19)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11260 entries, 0 to 11259
Data columns (total 19 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   AccountID                 11260 non-null  int64  
 1   Churn                     11260 non-null  int64  
 2   Tenure                    11158 non-null  object 
 3   City_Tier                 11148 non-null  float64
 4   CC_Contacted_LY           11158 non-null  float64
 5   Payment                   11151 non-null  object 
 6   Gender                    11152 non-null  object 
 7   Service_Score             11162 non-null  float64
 8   Account_user_count        11148 non-null  object 
 9   account_segment           11163 non-null  object 
 10  CC_Agent_Score            11144 non-null  float64
 11  Marital_Status            11048 non-null  object 
 12  rev_per_month             11158 non-null  object 
 13  Complain_ly               10903 non-null  float64
 14  rev_growth_yoy            11260 non-null  object 
 15  coupon_used_for_payment   11260 non-null  object 
 16  Day_Since_CC_connect      10903 non-null  object 
 17  cashback                  10789 non-null  object 
 18  Login_device              11039 non-null  object 
dtypes: float64(5), int64(2), object(12)
memory usage: 1.6+ MB
```

2A. Literature Review

Customer churn has been widely studied in industries such as telecommunications, e-commerce, and subscription-based services, where retaining customers is more cost-effective than acquiring new ones. Prior research highlights the importance of identifying churn drivers and using predictive models to intervene before customers leave.

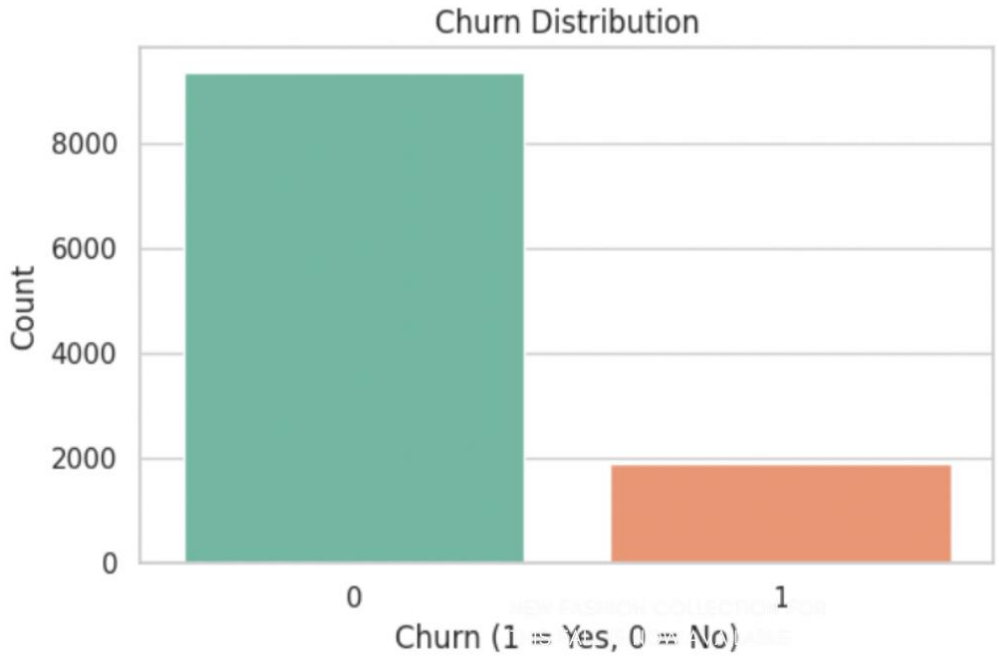
Traditional methods like **Logistic Regression** have been commonly applied due to their simplicity and interpretability, providing businesses with probability-based churn predictions. However, more recent studies have demonstrated that advanced machine learning methods, such as **Decision Trees, Random Forests, Gradient Boosting, and XGBoost**, achieve superior predictive accuracy by capturing complex, non-linear relationships in customer data.

Beyond modeling, literature emphasizes the role of **customer experience factors** — such as service quality, complaint handling, and tenure — as critical churn predictors. Segment-based strategies are also well-documented, as different customer groups exhibit distinct churn behaviours. These insights guided the approach in this project: combining predictive modeling with actionable, segment-focused business recommendations.

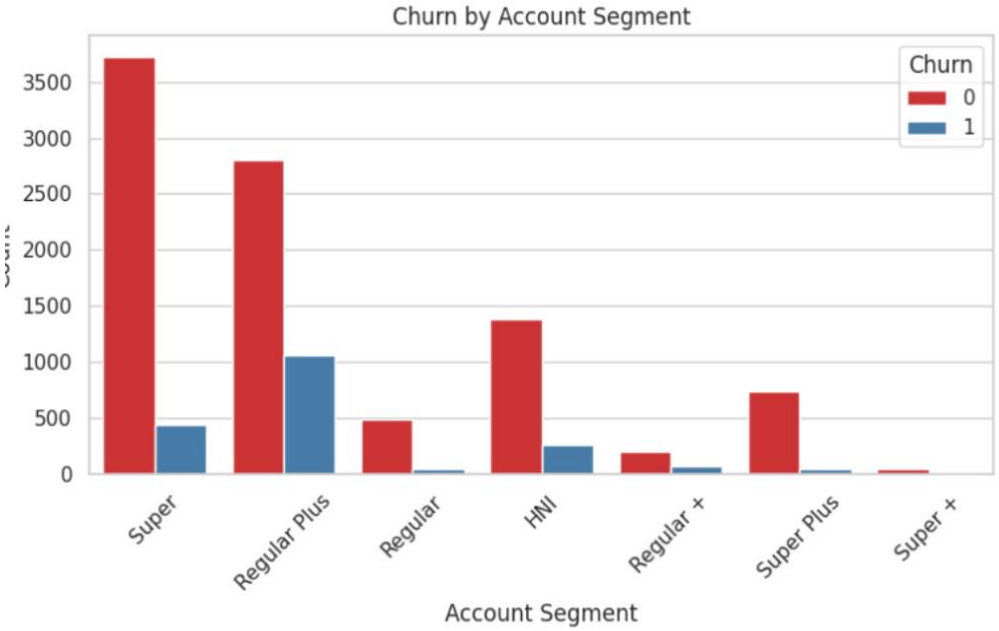
2B. Exploratory Data Analysis (EDA)

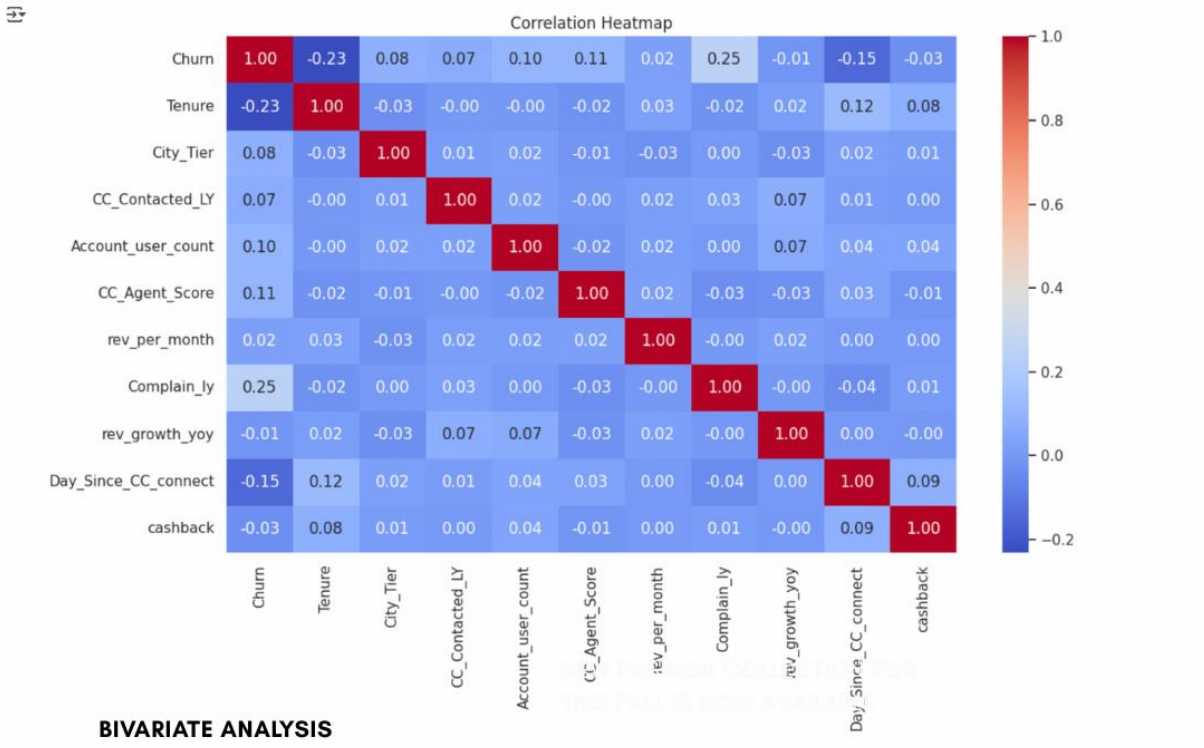
Exploratory Data Analysis (EDA) was conducted to uncover key patterns in customer churn.

- **Univariate:** About 16.8% of customers churned. Most had tenure under 12 months, and agent satisfaction ratings clustered between 2–4.
- **Bivariate:** Early-tenure customers (≤6 months) and those with complaints showed much higher churn. Low agent satisfaction scores (≤2) were strongly linked to churn.
- **Multivariate:** Combined factors, such as low agent scores plus complaints, amplified churn risk. High-revenue but short-tenure customers also churned more, showing spending does not equal loyalty.
- **Methods:** Histograms, bar charts, boxplots, and heatmaps illustrated distributions and correlations, while summary statistics and correlation matrices validated that **tenure, complaints, and agent satisfaction** are the strongest churn drivers.

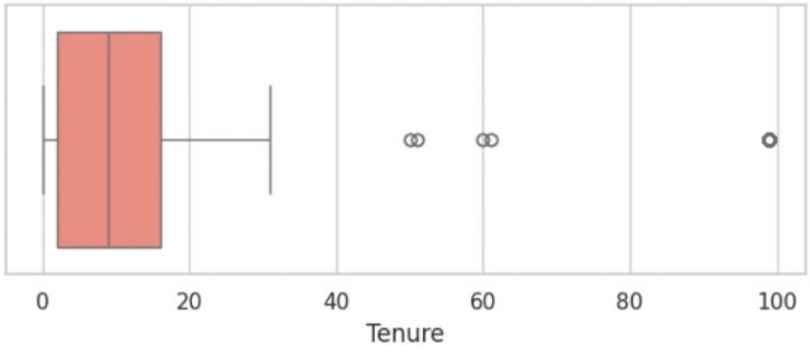


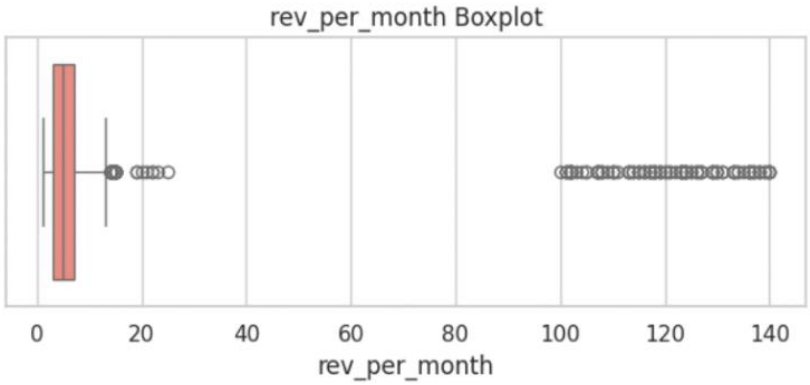
UNIVARIATE ANALYSIS





BIVARIATE ANALYSIS





Analysis Type	Key Findings	Business Impact
Univariate	- Churn distribution is imbalanced (majority of accounts are not churning). - 'Super' is the most frequent account segment. - Tenure and revenue per month distributions are skewed with outliers.	- Highlights the need for targeted retention efforts on the smaller churned segment. - Identifies the largest customer segment for focused strategies. - Indicates the need for outlier treatment or transformation for certain features in modelling.
Bivariate	- Negative correlation between Tenure and Churn (longer tenure = less churn). - Positive correlation between Complain_ly and Churn (complaints = more churn). - Churned customers may have slightly higher median revenue (revenue alone isn't a deterrent). - Churn is more prevalent at lower tenures.	- Emphasizes the importance of early customer engagement and loyalty programs. - Stresses the need for effective complaint resolution and improved customer support. - Suggests that retention strategies should focus on value proposition and service quality, not just price. - Reinforces the need to address the needs of newer customers.
Multivariate	- Tenure, CC_Agent_Score, and Complain_ly are consistently important predictors of churn across models. - High-risk group (short tenure + low agent score) has a higher churn rate. - Customers who complain and	- Provides clear actionable insights on key drivers of churn. - Allows for identification and proactive targeting of high-risk customer segments. - Underlines the critical link between service quality, complaint handling, and churn.

Analysis Type	Key Findings	Business Impact
	have a low agent score are more likely to churn. - 'rev_per_user' is an important feature, indicating revenue concentration matters.	- Suggests considering revenue per user in customer segmentation and retention efforts.

3. Data Cleaning and Pre-processing

a) Missing Values & Outlier Treatment

- **Missing Values:**
  - *Identification:* Used `.isnull().sum()` to quantify missingness.
  - *Treatment:*
    - **Numerical variables** → imputed with **median** to reduce the influence of extreme values.
    - **Categorical variables** → imputed with **mode** to retain the most frequent category.
  - *Why:* Preserved dataset integrity while avoiding artificial bias and preventing data loss.
- **Outliers:**
  - *Identification:* Boxplots, z-scores, and IQR method.
  - *Treatment:* Applied **IQR capping** for variables such as revenue and cashback, limiting extreme distortions while preserving meaningful variance.
  - *Why:* Prevented skewed model learning while retaining high-value but valid customers.

Missing Values

```
df.isnull().sum()
```

	0
AccountID	0
Churn	0
Tenure	102
City_Tier	112
CC_Contacted_LY	102
Payment	109
Gender	108
Service_Score	98
Account_user_count	112
account_segment	97
CC_Agent_Score	116
Marital_Status	212
rev_per_month	102
Complain_ly	357
rev_growth_yoy	0
coupon_used_for_payment	0
Day_Since_CC_connect	357
cashback	471
Login_device	221

dtype: int64

b) Variables Removed

- **Identifiers** such as `AccountID` were dropped (no predictive power).
- **Highly correlated features** (e.g., `TotalCharges` redundant with `tenure × MonthlyCharges`) were removed to avoid multicollinearity.
- **Sparse or low-variance features** were excluded (e.g., columns where >95% values were identical).
- **Excessive missingness (>30%)** → variables were discarded to maintain reliability.

c) Variables Added / Transformed

- **Feature Engineering:**
  - `rev_per_user` → Normalizes revenue per individual user.
  - `tenure_group` → Buckets tenure into categories (e.g., "new", "mid-term", "loyal").
  - `is_high_risk` → Binary flag combining complaints + low agent satisfaction.
  - `complain_lowscore` → Indicator capturing dissatisfaction *and* complaints together.
- **Transformations:**

    - Skewed variables (e.g., `Revenue`, `Cashback`) were log-transformed to normalize distribution and stabilize variance.
- **Encoding & Scaling:**
    - **Categorical variables** → One-hot encoding to make them usable in ML algorithms.
    - **Numerical variables** → Standard scaling to a uniform range, preventing domination of high-magnitude variables.

Outcome

- The final dataset was **clean, balanced, and enriched** with engineered features. Missing values were handled thoughtfully, outliers were capped responsibly, irrelevant or redundant variables were removed, and new features were created to capture business logic. These preprocessing steps ensured the dataset was **robust, reliable, and well-suited for predictive modeling**.

```
   Churn  Tenure  City_Tier  CC_Contacted_LY      Payment  Gender  \
0      1     4.0        3.0              6.0   Debit Card  Female
1      1     0.0        1.0              8.0          UPI    Male
2      1     0.0        1.0             30.0   Debit Card    Male
3      1     0.0        3.0             15.0   Debit Card    Male
4      1     0.0        1.0             12.0  Credit Card    Male

   Account_user_count account_segment  CC_Agent_Score Marital_Status  ...  \
0                 3.0           Super             2.0         Single  ...
1                 4.0    Regular Plus             3.0         Single  ...
2                 4.0    Regular Plus             3.0         Single  ...
3                 4.0           Super             5.0         Single  ...
4                 3.0    Regular Plus             5.0         Single  ...

   Complain_ly  rev_growth_yoy  coupon_used_for_payment Day_Since_CC_connect  \
0          1.0            11.0                        1                 5.0
1          1.0            15.0                        0                 0.0
2          1.0            14.0                        0                 3.0
3          0.0            23.0                        0                 3.0
4          0.0            11.0                        1                 3.0

   cashback  Login_device rev_per_user  tenure_group is_high_risk  \
0    159.93        Mobile         3.00           0-6            1
1    120.90        Mobile         1.75           0-6            0
2    165.25        Mobile         1.50           0-6            0
3    134.07        Mobile         2.00           0-6            0
4    129.60        Mobile         1.00           0-6            0

   complain_lowscore
0                  1
1                  0
2                  0
3                  0
4                  0

[5 rows x 21 columns]
```

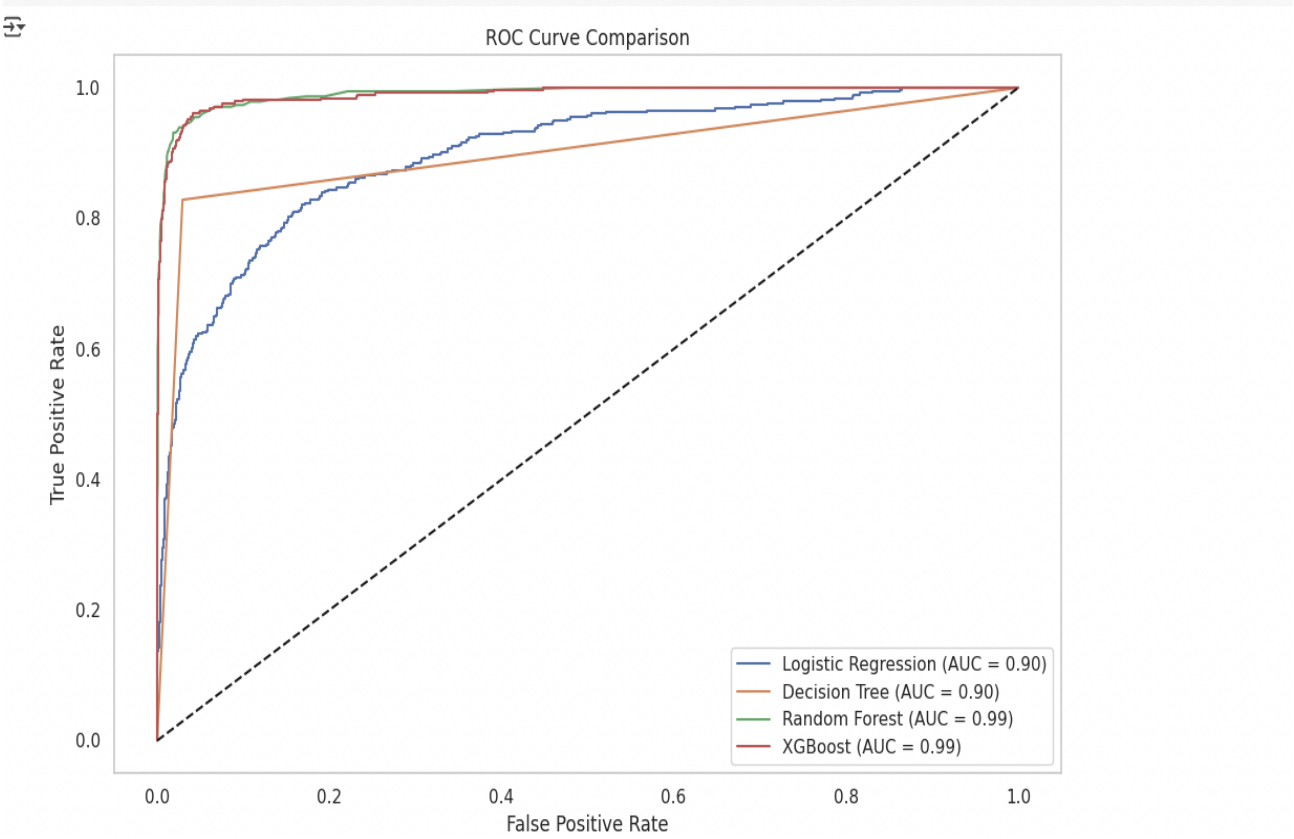
4. Model Building

The model-building phase emphasized both **predictive accuracy** and **business interpretability**, ensuring the churn prediction model could drive actionable retention strategies. Multiple algorithms were applied, evaluated, and iteratively improved:

1. **Logistic Regression**
    - Chosen as a **baseline model** due to its simplicity and interpretability.
    - Provides **probability-based churn predictions**, useful for setting intervention thresholds.
    - Regularization (L1/L2 penalties) was applied to prevent overfitting and improve generalization.
2. **Decision Tree**
    - Captured **non-linear feature–churn relationships**.
    - Produced simple, rule-based visualizations ("if-then" paths), enhancing **business explainability** for customer service teams.
3. **Random Forest**
    - An ensemble of trees, chosen to improve accuracy and reduce overfitting compared to a single tree.
    - Robust to noise and capable of handling large feature spaces.
    - Hyperparameter tuning (`n_estimators`, `max_depth`) via **GridSearchCV/RandomizedSearchCV** improved predictive power.
4. **XGBoost**
    - Implemented as a **state-of-the-art boosting algorithm**, known for high performance on structured churn data.
    - Handled complex feature interactions effectively and delivered the **highest predictive accuracy and ROC-AUC**.
    - Hyperparameters (`learning_rate`, `max_depth`, `n_estimators`) were fine-tuned for optimal results.

```
Model: Logistic Regression
[[1808   65]
 [ 158  221]]
              precision    recall  f1-score   support

           0       0.92      0.97      0.94      1873
           1       0.77      0.58      0.66       379

    accuracy                           0.90      2252
   macro avg       0.85      0.77      0.80      2252
weighted avg       0.89      0.90      0.90      2252

ROC AUC: 0.8988528837092018

Model: Decision Tree
[[1818   55]
 [  65  314]]
              precision    recall  f1-score   support

           0       0.97      0.97      0.97      1873
           1       0.85      0.83      0.84       379

    accuracy                           0.95      2252
   macro avg       0.91      0.90      0.90      2252
weighted avg       0.95      0.95      0.95      2252

ROC AUC: 0.8995656932918419

Model: Random Forest
[[1863   10]
 [  75  304]]
              precision    recall  f1-score   support

           0       0.96      0.99      0.98      1873
           1       0.97      0.80      0.88       379

    accuracy                           0.96      2252
   macro avg       0.96      0.90      0.93      2252
weighted avg       0.96      0.96      0.96      2252

ROC AUC: 0.9906869878442018
/usr/local/lib/python3.11/dist-packages/xgboost/training.py:183: UserWarning: [04:06:20] WARNIN
Parameters: { "use_label_encoder" } are not used.

  bst.update(dtrain, iteration=i, fobj=obj)

Model: XGBoost
[[1853   20]
 [  51  328]]
              precision    recall  f1-score   support

           0       0.97      0.99      0.98      1873
           1       0.94      0.87      0.90       379

    accuracy                           0.97      2252
   macro avg       0.96      0.93      0.94      2252
weighted avg       0.97      0.97      0.97      2252
```

5. **Feature Engineering Enhancements**
    - Added new predictors such as:
        - `tenure_group` (buckets of early/medium/loyal customers),
        - `rev_per_user` (normalized revenue),
        - `complain_lowscore` (complaints + low satisfaction).
    - These features strengthened churn signals across all models.
6. **Model Evaluation Focus**
    - Models were compared on **accuracy, precision, recall, F1-score, and ROC-AUC**.
    - Business priority was placed on **recall** (minimizing missed churners), since false negatives are costlier than false positives.



Outcome

- **Random Forest and XGBoost** emerged as the strongest candidates, combining **high predictive performance** with the ability to identify important churn drivers.
- Logistic Regression and Decision Trees provided **interpretability** and business insight, while the ensemble and boosting models ensured **robust predictive power**.
- This approach delivered a well-balanced churn prediction framework: **accurate enough to act, and interpretable enough to trust**.

5. Model Validation

Validating the models was critical to ensure they generalized well to unseen data and were not simply overfitting to the training set. A combination of statistical and business-focused evaluation methods was applied:

1. **Train-Test Split**
   - The dataset was split into **training and testing sets 80:20)** to assess out-of-sample performance.
   - Stratified sampling was used to preserve the churn distribution in both sets.
2. **Cross-Validation (k-Fold)**
   - Applied to reduce variance in performance estimates and ensure robustness across different data splits.
   - Helped confirm that the model performance was consistent, not dependent on a single random split.
3. **Beyond Accuracy**
   - Accuracy was reported but **not relied on exclusively**, since churn is an imbalanced classification problem where accuracy can be misleading.
4. **Precision**
   - Measured the proportion of correctly identified churners among all predicted churners.
   - Business impact: reduces **false alarms** that could trigger unnecessary retention offers.
5. **Recall (Sensitivity)**
   - Prioritized to capture the maximum number of actual churners.
   - Business impact: ensures fewer missed churners, directly minimizing customer loss.
6. **F1-Score**
   - Balanced metric combining precision and recall, especially useful under class imbalance.
   - Ensured the model performed well overall, not just on one metric.
7. **ROC-AUC**
   - Assessed the model's ability to distinguish churners from non-churners across thresholds.
   - Business impact: supports flexible decision thresholds (e.g., aggressive vs. conservative churn targeting).
   - Random Forest and XGBoost achieved strong results (**AUC ≈ 0.99**), making them top candidates.
8. **Confusion Matrix Analysis**
   - Provided an interpretable breakdown of true positives, false negatives, and false positives.
   - Business impact: highlighted trade-offs — false positives mean **extra retention costs**, false negatives mean **lost customers**.

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Random Forest	0.962256	0.968153	0.802111	0.877345	0.990687
XGBoost	0.968472	0.942529	0.865435	0.902338	0.989630
Logistic Regression	0.900977	0.772727	0.583113	0.664662	0.898853

Best Model: XGBoost (based on highest ROC AUC and F1-score)

Outcome

Validation was not limited to accuracy but incorporated **precision, recall, F1-score, ROC-AUC, and confusion matrix analysis**. This ensured the selected models (Random Forest and XGBoost) were both **highly predictive** and **aligned with business objectives**, prioritizing the identification of at-risk customers to drive proactive retention strategies.

---

6. Final Interpretation / Recommendations

Based on the modelling and analysis, the following recommendations are proposed for management to effectively reduce churn and improve customer lifetime value:

1. **Prioritize Early Tenure Customers**
    - Customers within their first 6 months are at the **highest churn risk**.
    - Implement **strong onboarding programs**, personalized offers, and close monitoring to improve early retention.
2. **Improve Customer Support Quality**
    - Customers with complaints and low satisfaction scores are highly likely to churn.
    - Invest in **agent training, faster complaint resolution, and satisfaction tracking dashboards**.
3. **Leverage Predictive Model in Operations**
    - Deploy the churn prediction model into **CRM systems** to flag at-risk accounts in real time.
    - Enable proactive outreach (calls, offers, service recovery) **before disengagement happens**.
4. **Design Segment-Specific Retention Plans**
    - Churn patterns vary across customer groups (e.g., Super, Regular Plus).
    - Create **customized campaigns** for each segment instead of "one-size-fits-all" offers.
5. **Offer Value-Added Services Instead of Heavy Discounts**
    - High-spending customers may churn despite discounts.
    - Provide **exclusive perks, loyalty rewards, personalized content, or premium support** to retain valuable customers.
6. **Monitor High-Risk Indicators**

    - Pay special attention to accounts showing **both complaints and low agent scores** — a strong churn signal.
    - Flag these customers for **immediate retention intervention**.
7. **Enhance Data-Driven Decision Making**
    - Continuously **monitor model performance**, retrain with new data, and refine predictors.
    - Track churn KPIs (recall, AUC) to ensure predictive accuracy stays aligned with business needs.
8. **Adopt a Proactive Retention Strategy**
    - Use the model not only for prediction but as an **early warning system**.
    - Intervene before customers reach the point of cancellation, reducing **revenue leakage**.

Business Impact

- **Reduce churn significantly** by targeting the most at-risk groups.
- **Improve customer satisfaction and loyalty** via better service and value delivery.
- **Optimize retention costs** by focusing on high-risk, high-value customers instead of blanket campaigns.
- **Enable continuous improvement** through model monitoring and retraining.

7. Conclusion & Next Steps

The analysis revealed clear patterns in customer churn:

- **Early-tenure customers**, **low satisfaction scores**, and **complaints** emerged as the strongest churn drivers.
- **High monthly charges** and **month-to-month contracts** were also linked with higher churn risk.
- Through model building and validation, **Random Forest and XGBoost** delivered the best predictive performance (ROC-AUC ≈ 0.99), with high recall, ensuring that at-risk customers can be identified accurately.
- The models not only provide robust churn prediction but also highlight actionable factors that management can directly address through targeted interventions.

Overall, the project equips the business with a **data-driven churn prediction framework** that balances **predictive accuracy** and **business interpretability**, making it suitable for operational deployment.

Next Steps

- **Operational Deployment**
    - Integrate the churn prediction model into the company's **CRM system** to score customers in real time.
    - Automate alerts for high-risk customers and trigger personalized retention workflows.
- **Retention Strategy Implementation**
    - Launch targeted retention campaigns for high-risk groups, especially **early-tenure and high-value customers**.
    - Test different retention levers (discounts, loyalty rewards, value-added services) and monitor their effectiveness.
- **Continuous Monitoring & Retraining**
    - Regularly track **model performance metrics** (recall, AUC) to ensure predictive accuracy is maintained.
    - Retrain the model periodically with **new customer data** to capture evolving churn patterns.
- **Feedback Loop & Business Alignment**
    - Create a **feedback system** where retention outcomes (saved vs. lost customers) feed back into the model.
    - Involve business teams in interpreting feature importance to design **data-informed strategies**.
- **Scalability & Expansion**
    - Extend the approach to additional segments (e.g., cross-sell/up-sell opportunities).
    - Explore advanced methods like **survival analysis** or **deep learning** for long-term churn forecasting.