

## Customer Churn Analysis Project Summary

### Data Features and Preprocessing

Key Features Used:

Category	Features
Customer Behavior	Tenure, OrderCount, HourSpendOnApp, DaySinceLastOrder
Satisfaction Indicators	SatisfactionScore, Complain
Economic Factors	CashbackAmount, CouponUsed, OrderAmountHikeFromLastYear
Logistics	WarehouseToHome, CityTier
Technical Engagement	NumberOfDeviceRegistered, NumberOfAddress

### Data Cleaning:

- Converted columns to numeric formats
- Replaced zeros with Nan in CashbackAmount & CouponUsed, filled with missing values (medians for numeric, modes for categorical)
- Standardized features for modeling, removed zero-variance columns

### Prediction Models and Evaluation

**Target Variable:** Binary customer churn (0/1)

### Regression Analysis:

- Univariate & Multivariate Random Forest Regression ( $R^2 \approx 0.3$ )

### Clustering Analysis:

Clustering Method	Number of Clusters	Evaluation Metrics	Best Performing
Agglomerative Clustering	3	Silhouette Score, PCA Visualization, Churn Rate Differentiation	No
K-Means Clustering	3	Silhouette Score, PCA Visualization, Churn Rate Differentiation	No
Mini-Batch K-Means	3	Silhouette Score, PCA Visualization, Churn Rate Differentiation	No
Mean Shift Clustering	7 (auto determined)	Silhouette Score: 0.299	Yes (Best Performing)

Cluster-Wise Churn Rates

Cluster	Number of Customers	Churn Rate	Remarks
Cluster 2	41	31.71% (High Risk)	High Churn Segment
Cluster 0	5,505	16.89%	Moderate Churn Rate
Clusters 3, 4, 5	50	0%	Perfect Loyalty
Cluster 6	1	100%	Complete Churn

Classification Analysis in a tabular format:

Method	Accuracy	ROC AUC	F1 Score (Churned Class)	Cluster Feature Impact
Logistic Regression	N/A	N/A	N/A	Tested
K-Nearest Neighbors	N/A	N/A	N/A	Tested
Decision Tree (Best Model)	94.2%	0.901	0.825	Improved Performance
Support Vector Machine	N/A	N/A	N/A	Tested
Naive Bayes	N/A	N/A	N/A	Tested
Neural Network (Runner-up)	94.1%	0.971 (Best Discrimination)	0.806	Improved Performance

Evaluation Metrics Considered:

- Accuracy, ROC AUC, F1 Score (for churn class)
- Cross-validation scores, confusion matrices, comparison with and without cluster features

Conclusion & Key Takeaways

- Customer Segmentation: Mean Shift revealed natural churn patterns, outperforming arbitrary cluster sizes.
- Classification Effectiveness: Decision Trees balanced accuracy & interpretability, making them ideal for business use.
- Cluster Feature Impact: Improved classification performance, especially for Neural Networks, validating a two-stage approach.
- Business Application:
  - Target retention efforts on high-risk Cluster 2 (31.7% churn)
  - Study zero-churn clusters to identify loyalty drivers
  - Implement segment-specific retention strategies
- Methodology Insight: The strong performance of Mean Shift **underscores the importance of data-driven clustering over predefined counts.**