**Customer Churn Analysis Project Summary**

**Data Features and Preprocessing**

**Key Features Used:**

* Customer behavior metrics: Tenure, OrderCount, HourSpendOnApp, DaySinceLastOrder
* Satisfaction indicators: SatisfactionScore, Complain
* Economic factors: CashbackAmount, CouponUsed, OrderAmountHikeFromlastYear
* Logistics metrics: WarehouseToHome, CityTier
* Technical engagement: NumberOfDeviceRegistered, NumberOfAddress

**Data Cleaning:**

* Converted columns to appropriate numeric formats
* Replaced zeros with NaN in CashbackAmount and CouponUsed
* Filled missing numeric values with medians
* Filled missing categorical values with modes
* Standardized features for modeling
* Dropped zero-variance columns

**Prediction Models and Evaluation**

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

**Initial Regression Analysis:**

* Univariate Random Forest Regression for each feature
* Multivariate Random Forest Regression with R² of approximately 0.3

**Clustering Analysis**

**Methods Evaluated:**

1. Agglomerative Clustering (3 clusters)
2. K-Means Clustering (3 clusters)
3. Mini-Batch K-Means (3 clusters)
4. Mean Shift Clustering (7 clusters, automatically determined)

**Evaluation Criteria:**

* Silhouette score (measure of cluster coherence)
* PCA visualization of cluster separation
* Churn rate differentiation between clusters

**Best Performing Method:** Mean Shift with silhouette score of 0.299

* Generated 7 natural clusters with distinct churn behaviors:
  + Cluster 0 (5,505 customers): 16.89% churn rate
  + Cluster 2 (41 customers): 31.71% churn rate (high-risk segment)
  + Clusters 3, 4, 5 (50 customers): 0% churn rate (perfect loyalty)
  + Cluster 6 (1 customer): 100% churn rate

**Classification Analysis**

**Methods Tested:**

1. Logistic Regression
2. K-Nearest Neighbors
3. Decision Tree
4. Support Vector Machine
5. Naive Bayes
6. Neural Network

**Evaluation Metrics:**

* Accuracy, ROC AUC, F1 score (particularly for minority churn class)
* Cross-validation scores
* Confusion matrices
* Each model was tested both with and without cluster features

**Best Classification Method:** Decision Tree with cluster feature

* Accuracy: 94.2%
* ROC AUC: 0.901
* F1 Score (Churned class): 0.825

**Runner-up:** Neural Network with cluster feature

* Accuracy: 94.1%
* ROC AUC: 0.971 (highest discrimination ability)
* F1 Score (Churned): 0.806

**Conclusion and Key Takeaways**

1. **Customer Segmentation Value:** Mean Shift clustering identified natural customer segments with significantly different churn behaviors, providing more business-relevant segmentation than forced equal-sized clusters.
2. **Classification Effectiveness:** Decision Trees offered the optimal balance between prediction accuracy and interpretability, crucial for business implementation.
3. **Cluster Feature Impact:** Including cluster assignments as features improved model performance, particularly for Neural Networks, validating the two-stage approach.
4. **Business Application:** The project provides actionable intelligence for:
   * Targeting retention efforts toward high-risk Cluster 2 (31.7% churn)
   * Studying zero-churn clusters to identify loyalty drivers
   * Implementing segment-specific retention strategies
5. **Methodology Insight:** The superior performance of Mean Shift clustering demonstrates the value of letting the data reveal its natural structure rather than imposing predetermined cluster counts.