ANA 630: Advanced Analytic Applications

CHURN ANALYSIS USING CLASSIFICATION AND CLUSTERING TECHNIQUES

BY WILLIAM OWENS

https://youtu.be/9pE9-ysKe1Y

What is Churn?

- Customer cancels their subscription with a service provider
- Churned customers equal lost revenue and profit:
  - $100 billion loss globally!*
- Costs more to acquire new customers than to retain existing customers:
  - 5-10 times more!*
- Reducing churn can save costs and increase revenue/profit:
  - 5% reduction = 25-85% revenue boost!*

*(Almana, Aksoy, & Alzahran, 2014)

Questions to Consider

What factors influence churn?

What types of customers churn?

Can we predict who will churn?

Research Outline

Statistical Analysis

Clustering Analysis

Predictive Analysis

Statistical Analysis

- Chi Square Tests
- CA Trend Tests
- Boxplot Analysis
- Logistic Regression
- Odds Ratios

Chi Square Tests

- Statistically significant association between complaint and churn
- Statistically significant association between tariff plan and churn
- Statistically significant association between active status and churn

Contingency Table for tariff_plan by churn:

tariff_plan	0	1
churn		
0	2416	239
1	489	6

Chi2 statistic: 34.22, p-value: 0.0000

Contingency Table for complaint by churn:

complaint	0	1
churn		
0	2614	41
1	295	200

Chi2 statistic: 886.21, p-value: 0.0000

Contingency Table for active_status by churn:

active_status	0	1
churn		
0	412	2243
1	370	125

Chi2 statistic: 781.11, p-value: 0.0000

Cochran-Armitage Trend Tests

- No trend exists for age group with respect to churn
- Trend exists for charge amount with respect to churn
- Age group may not be informative in later modeling

```
Results for Variable: age_group
Contingency Table:
```

age_group	1	2	3	4	5
churn					
0	123	853	1195	316	168
1	0	184	230	79	2

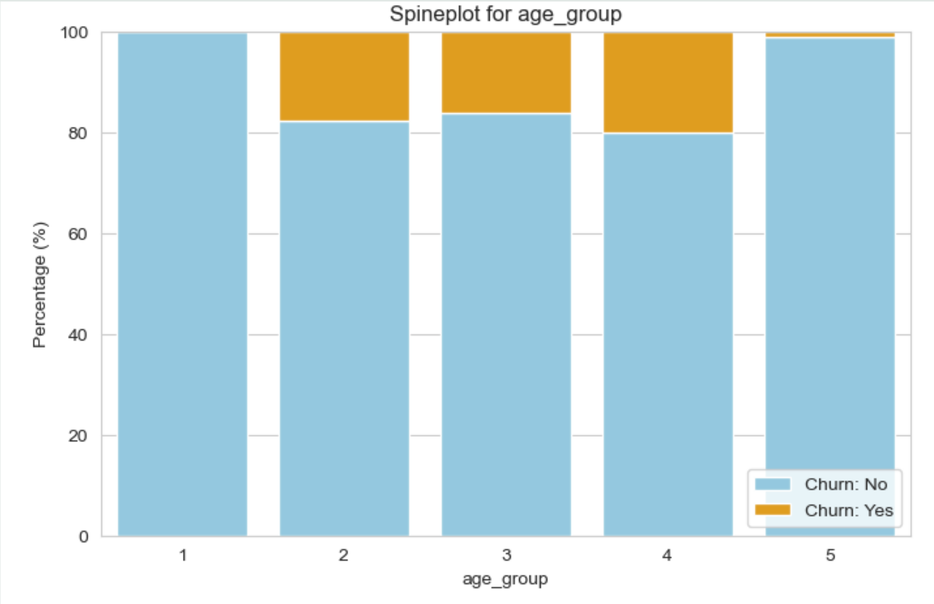
```
Statistic: 7518.0000
Null Mean: 7502.3449
Null SD: 18.2435
Z-score: 0.8581
P-value: 0.3908
```

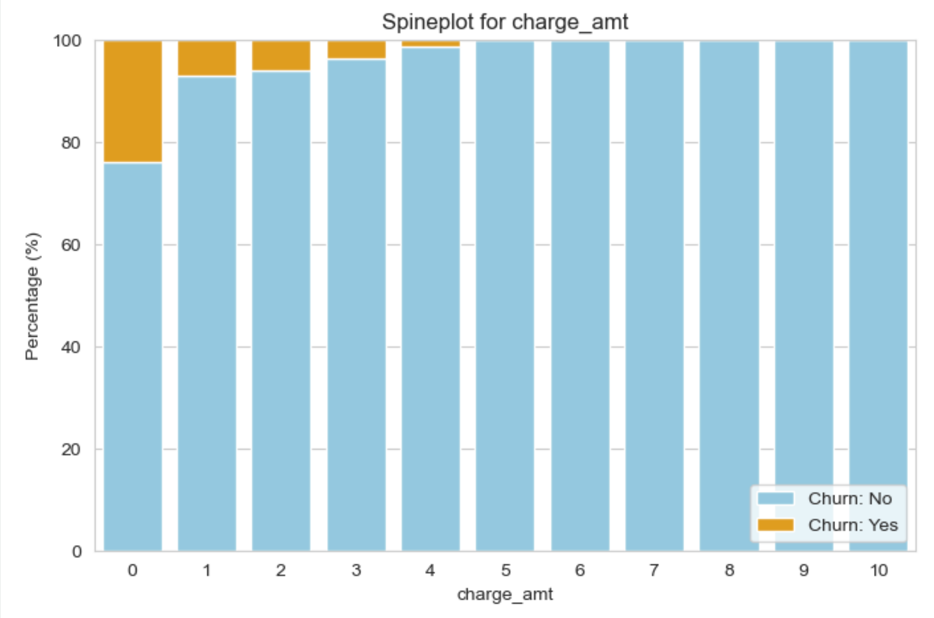
```
Results for Variable: charge_amt
Contingency Table:
```

charge_amt	0	1	2	3	4	5	6	7	8	9	10
churn											
0	1347	574	372	192	75	30	11	14	19	14	7
1	421	43	23	7	1	0	0	0	0	0	0

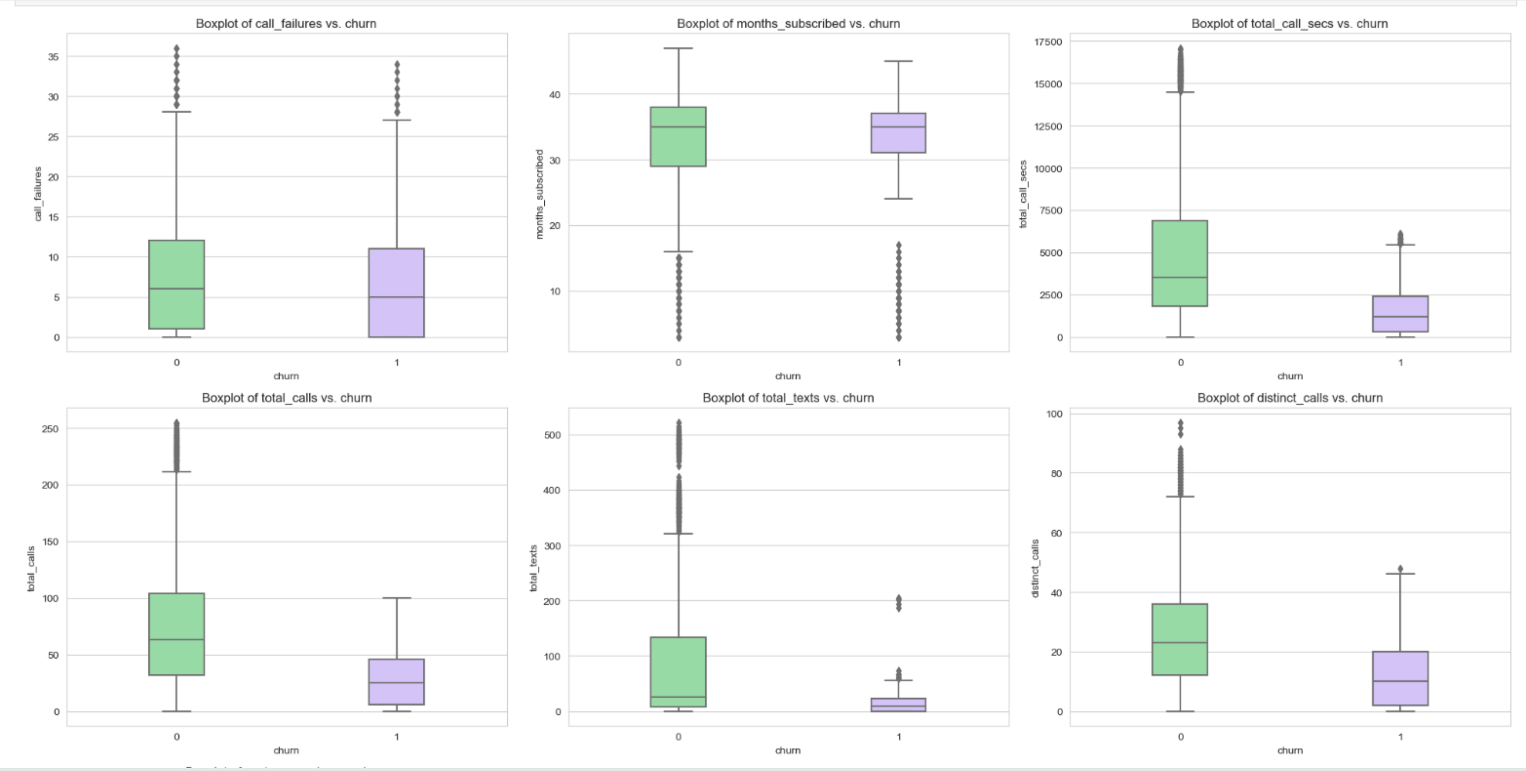
```
Statistic: 5511.0000
Null Mean: 5174.8501
Null SD: 31.4262
Z-score: 10.6965
P-value: 0.0000
```

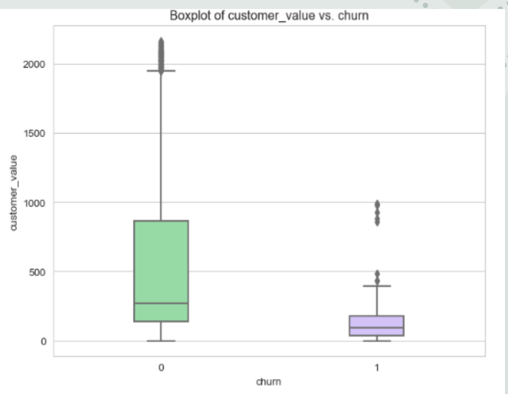
Spineplots


Spineplot for age_group
Percentage (%)
100
80
60
40
20
0
1
2
3
4
5
age_group
Churn: No
Churn: Yes


Spineplot for charge_amt
Percentage (%)
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
10
charge_amt
Churn: No
Churn: Yes

Boxplot Analysis





Lots of outliers!

Logistic Regression

- Initial Model:
  - Statistically insignificant variables: total call seconds, distinct calls, age group, and tariff plan
  - AIC = 1,411.13
  - BIC = 1,489.84
  - Deviance = 1,385.13
  - HL-Test = 22.90; p-value = 0.003
  - VIF over 10 = customer value, total texts, total calls, total call seconds, age group, months subscribed

Logistic Regression

- Reduced Model after Backward Elimination:
  - AIC = 1,407.87
  - BIC = 1,462.37
  - Deviance = 1,389.87
  - HL-Test = 28.74; p-value = 0.00035
  - VIF over 10 = customer value, total texts, total calls

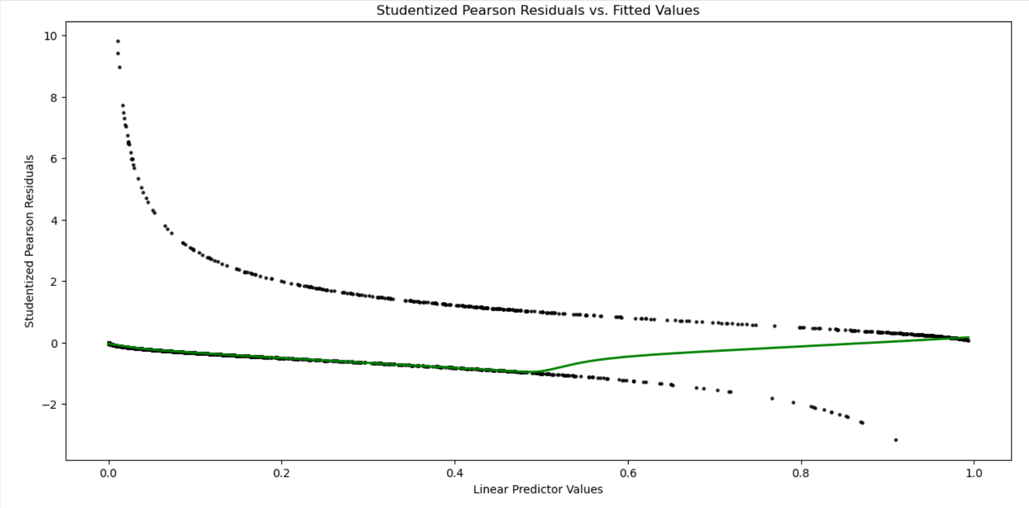
Logistic Regression

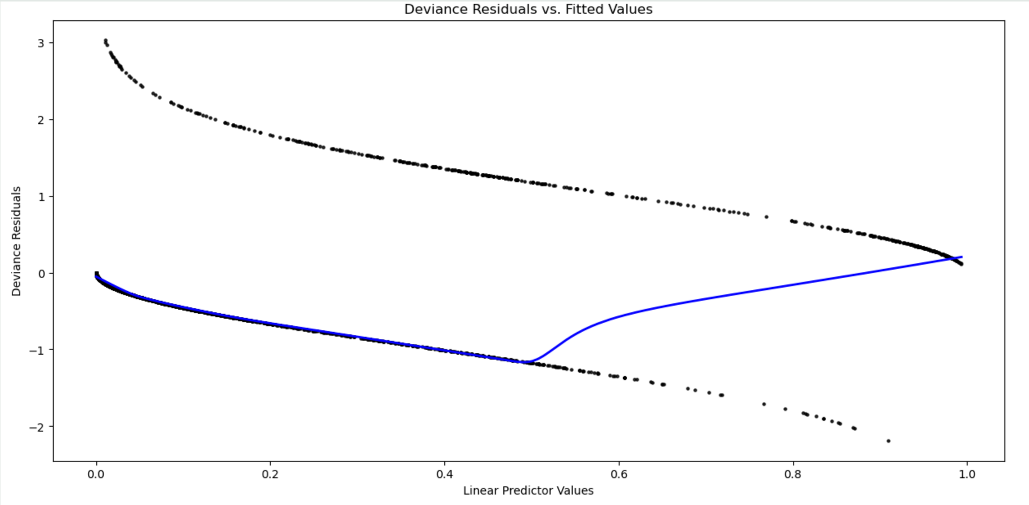
- Reduced Model with Interaction Term:
  - Interaction = complaint : active status
  - AIC = 1,384.35
  - BIC = 1,444.90
  - Deviance = 1,364.35
  - HL-Test = 12.49; p-value = 0.13
  - VIF over 10 = customer value, total texts, total calls

Adjusted Odds Ratios

- Complaint:
  - Point estimate = 19.95
  - 95% CI = 11.13 to 35.77
- Active status:
  - Point estimate = 0.23
  - 95% CI = 0.16 to 0.35

Residual Plots





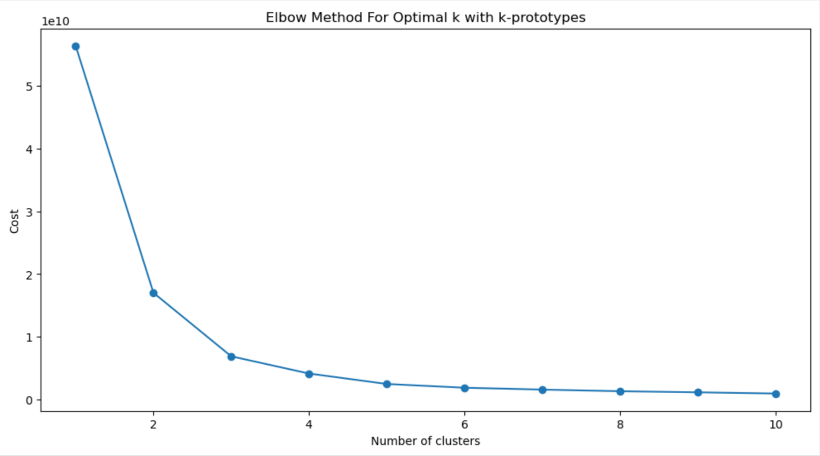
Cluster Analysis

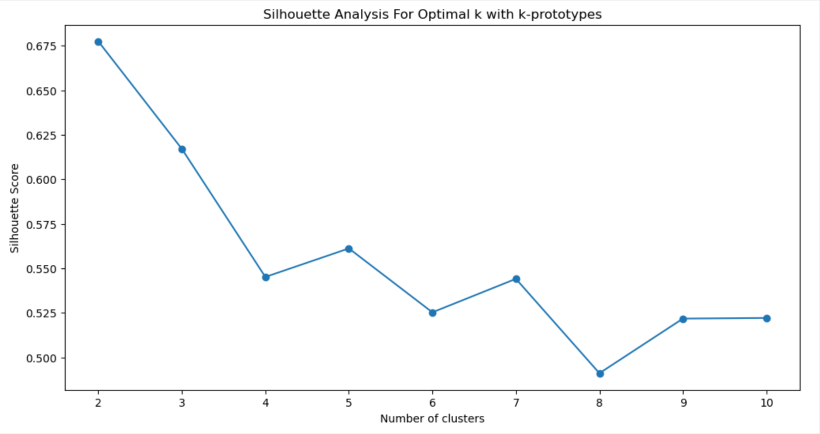
K-Prototypes

Factor Analysis of Mixed Data with K-Means

K-Prototypes

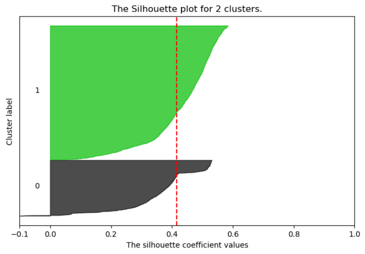
- Combo of K-Means and K-Modes
- K = 2, maybe 3
- 3 clusters may be more useful for segmentation
- Results:
  - Very High activity group
  - Moderate-to-High activity group
  - Low activity group

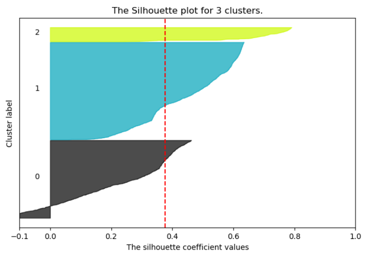


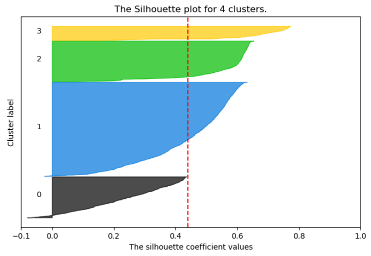


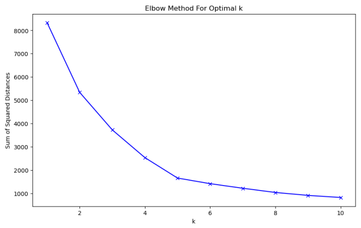
FAMD with K-Means

- Dimensionality reduction for mixed data
- K = 2, maybe 3
- 3 clusters used again
- Same results:
  - Very High activity group
  - Moderate-to-High activity group
  - Low activity group



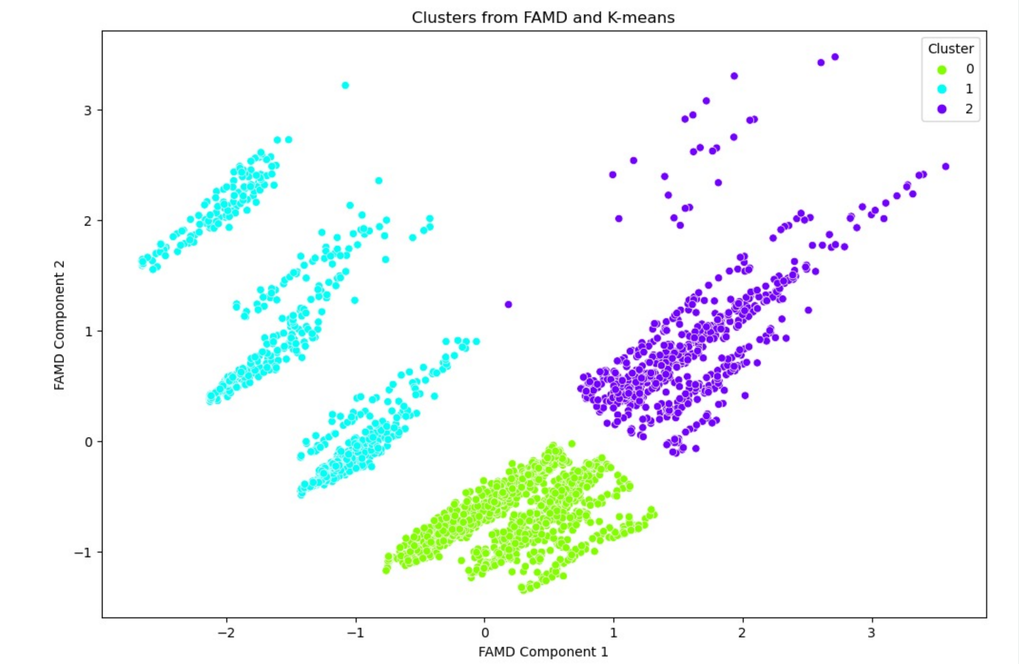






FAMD Plot

- Good separation
- Potentially more clusters out of cluster 2

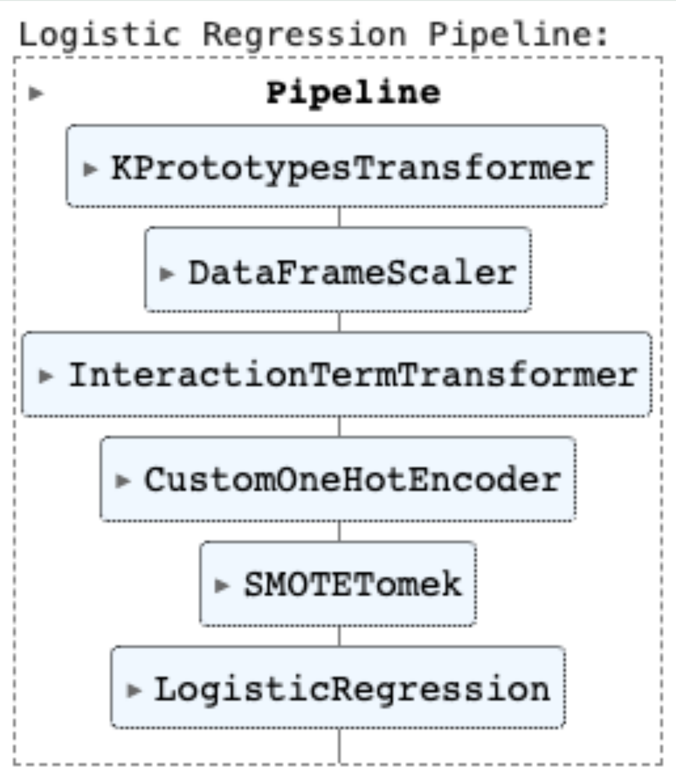


Predictive Analysis

- Preprocessing Workflow
- Classification Methods
- Hyperparameter Tuning
- Model Selection
- Final Model Evaluation
- Feature Importance

Preprocessing Workflow

- Training (50%), Validation (30%), and Test (20%) – stratified partitions
- Add cluster label feature using K-Prototypes
- Standardize continuous numeric features only
- Add interaction term feature (complaint : active status)
- One-hot encode cluster label feature
- Apply SMOTETomek for class balancing



Classification Methods

- Logistic Regression
- Random Forest
- Support Vector Machine
- XGBoost
- Histogram-based Gradient Boosting
- CatBoost
- Symbolic GP Classifer

Hyperparameter Tuning

- Randomized search algorithm
- Stratified 5-fold cross validation
- 250 iterations
- Best estimator = best mean recall score
- Fit best estimator to full training set

Model Selection

- Evaluated on validation set
- Top two models:
  - Histogram-based GB
  - XGBoost

Metrics for Logistic Regression:

	accuracy	recall	precision	f1_score	roc_auc
0	0.684656	0.858108	0.314356	0.460145	0.823722

Metrics for Random Forest:

	accuracy	recall	precision	f1_score	roc_auc
0	0.839153	0.905405	0.492647	0.638095	0.935675

Metrics for SVM:

	accuracy	recall	precision	f1_score	roc_auc
0	0.437037	0.993243	0.216814	0.355932	0.803982

Metrics for XGBoost:

	accuracy	recall	precision	f1_score	roc_auc
0	0.912169	0.824324	0.681564	0.746177	0.953241

Metrics for HistGB:

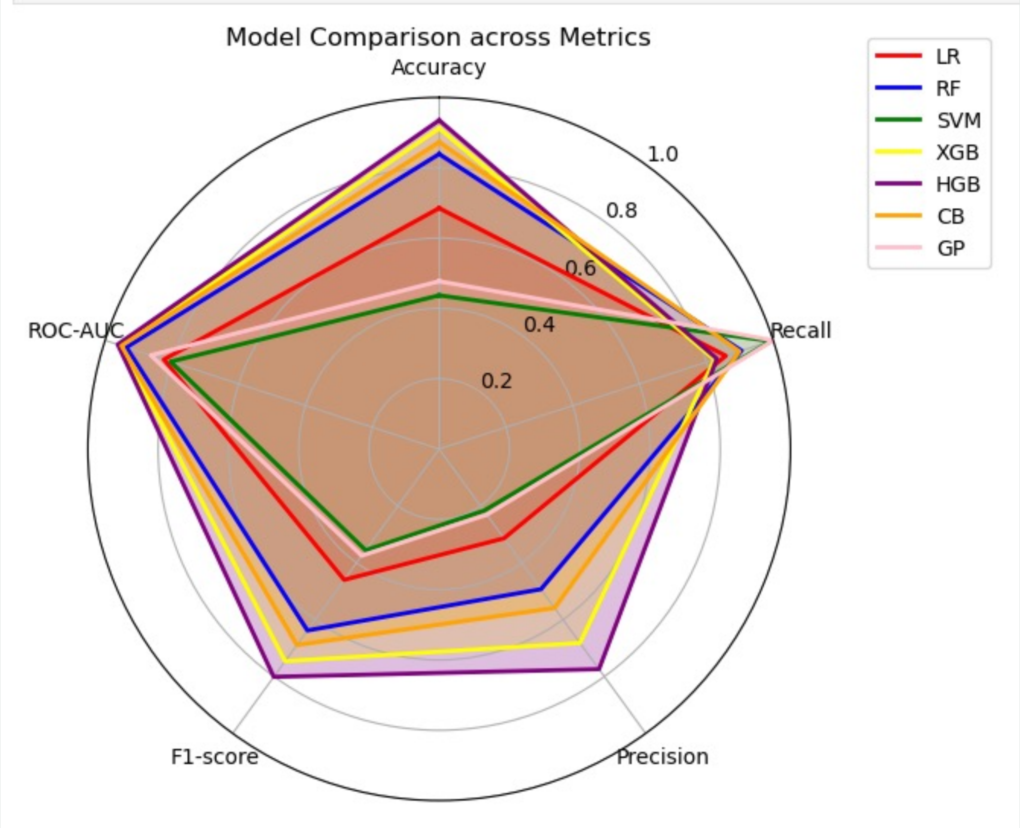
	accuracy	recall	precision	f1_score	roc_auc
0	0.93545	0.831081	0.773585	0.801303	0.962028

Metrics for CatBoost:

	accuracy	recall	precision	f1_score	roc_auc
0	0.873016	0.898649	0.558824	0.689119	0.95144

Metrics for Genetic Programming:

	accuracy	recall	precision	f1_score	roc_auc
0	0.477249	1.0	0.23053	0.374684	0.863428

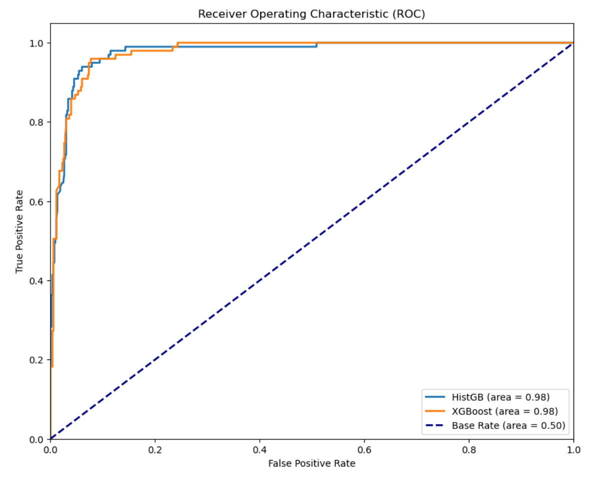


Final Model Evaluation

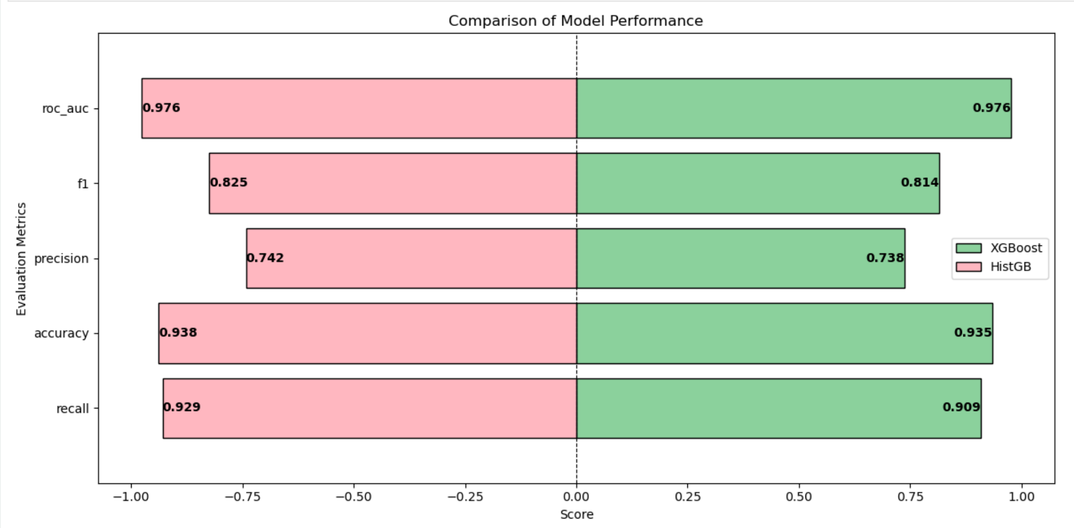
- Combined training and validation sets
- Retrained top two models on combined set
- Evaluated on test set
- Best model = Histogram-based GB (narrowly)
- Best recall + satisfactory precision
- Computationally efficient

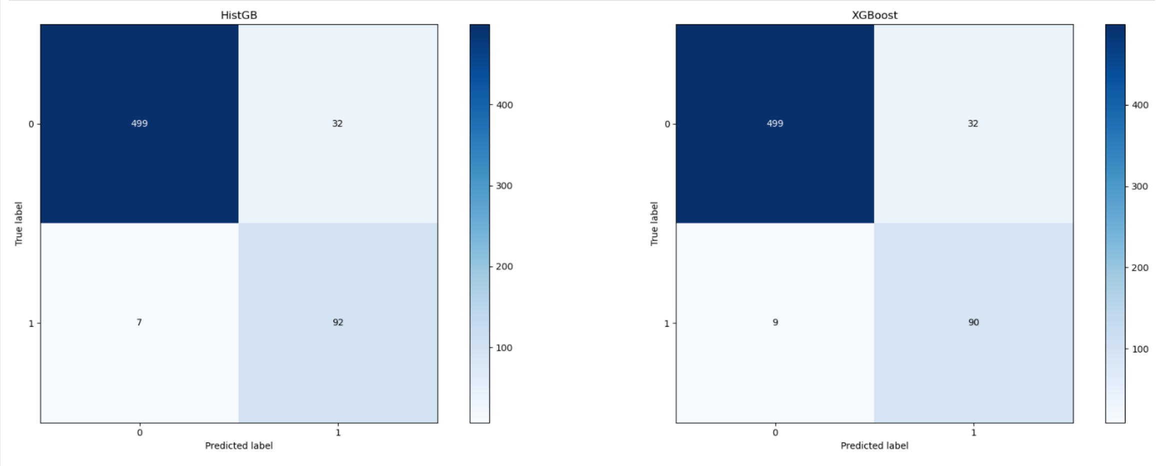
```
Test Results for HistGB:
accuracy: 0.9380952380952381
recall: 0.9292929292929293
precision: 0.7419354838709677
f1: 0.8251121076233184
roc_auc: 0.9763453746504593

Test Results for XGBoost:
accuracy: 0.9349206349206349
recall: 0.9090909090909091
precision: 0.7377049180327869
f1: 0.8144796380090498
roc_auc: 0.9764880442846545
```



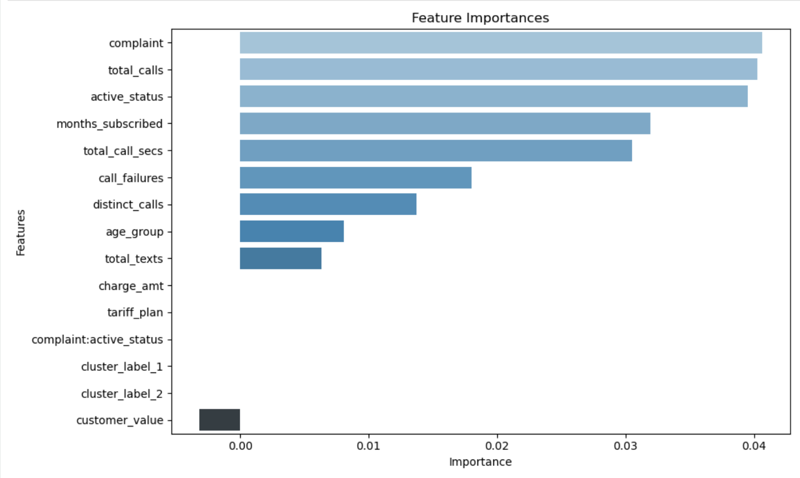
Comparison Plots

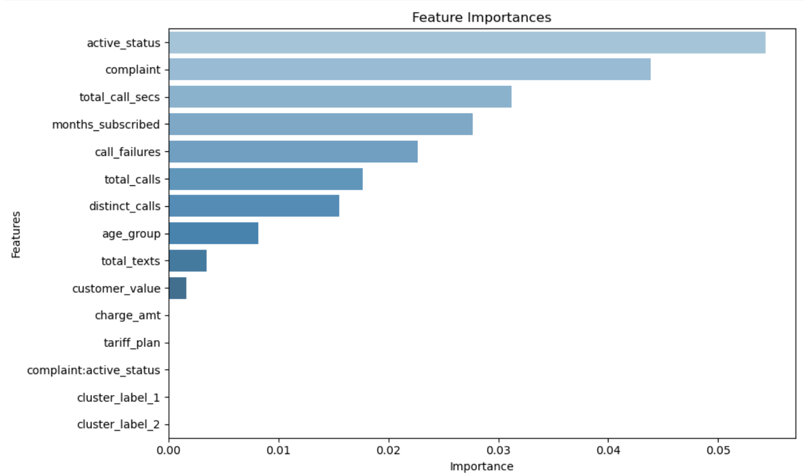




Feature Importance

- Important features = complaint, active status
- Non-important = clusters labels and interaction
- Redundancy of engineered features





Summary of Findings

- Complaint and active status appear significant
- Age may not be a factor for churn
- Strategies to address or preempt common complaints
- Strategies to boost customer activity
- More customer data to improve segmentation efforts
- Best predictive method = Histogram-based GB

References

- Ahn, J.-H., Han, S.-P., & Lee, Y.-S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications Policy, 30(10-11), 552-568. https://doi.org/10.1016/j.telpol.2006.09.006
- Almana, A. M., Aksoy, M. S., & Alzahran, R. (2014). A Survey On Data Mining Techniques In Customer Churn Analysis For Telecom Industry. International Journal of Engineering Research and Applications, 4(5), 165-171.
- Celik, O., & Osmanoglu, U. O. (2019). Comparing to Techniques Used in Customer Churn Analysis. Journal of Multidisciplinary Developments, 4(1), 30-38.
- Keramati, A., & Ardabili, S. M. S. (2011). Churn analysis for an Iranian mobile operator. Telecommunications Policy, 35(4), 344-356. https://doi.org/10.1016/j.telpol.2011.02.009

Thank You!