Customer Churn Analysis and Prediction

William Chen
Margaret Ji
Catherine Ruan

Overview

Goals	
Goals	▪ Use the consumer dataset to: ▪ Segment the Fulton Bank customer base ▪ Build a model that predicts customer churn

Agenda

1	Data processing
2	Customer segmentation
3	Feature scoring and predictive model implementation
4	Business recommendations

Data Processing

Objective

- Prepare data for analysis by removing and modifying data

Numerical/Binary

- Keep columns containing relevant characteristics of customer segments

Categorical

- Find appropriate level of detail
- One-hot encode

Balances

- Set missing balances to -10,000
- Use smooth symmetric log scaling

Missing Totals

- Fill blank cells with 0's or -1's depending on context

125 Columns

Customer Segmentation

Objective

- Figure out if consumers naturally fall into certain groups

Methods

- Dimensionality reduction
- Finding the optimal number of segments
- Clustering
- Segment analysis

Results

Churn

Behaviors

Customer Segmentation

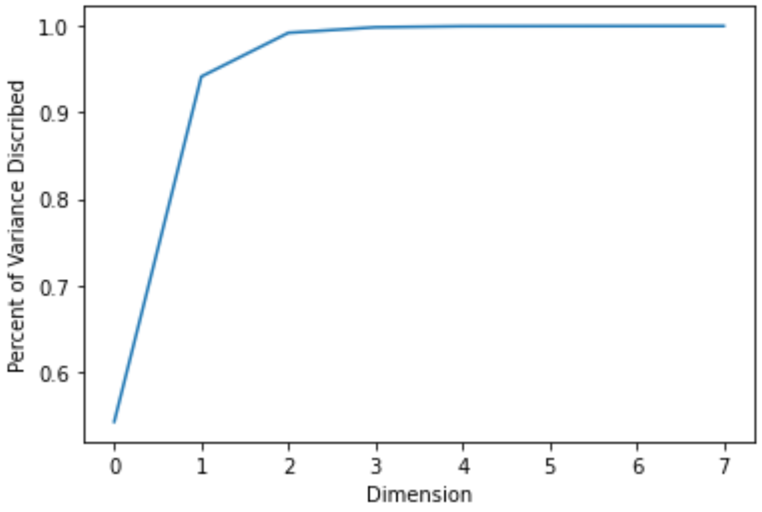
Objective

- Find a more concise representation of data

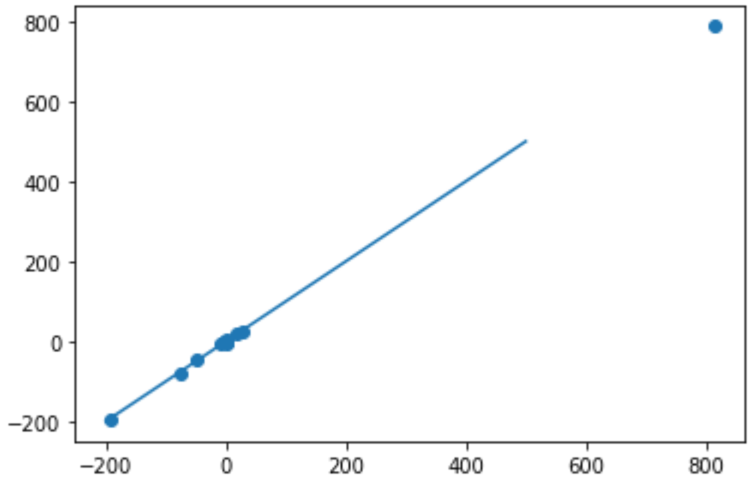
Method

- Dimensionality reduction
  - Autoencoder
  - Principal Component Analysis
- Performance analysis
  - Reconstruction loss

Percent of Data Described



Reconstructed Data vs. Actual Data



Customer Segmentation

Objective

- Use unsupervised learning to segment customers into groups

Why unsupervised segmentation?

- Cherry picking metrics may not capture nuances in the data
- Unsupervised clustering can cover as much information as possible
- To be understand churn, it is good to first understand its correlation with consumer behavior
- Spectral clustering is best suited for nonconvex geometry

Select best number of clusters based on eigengaps

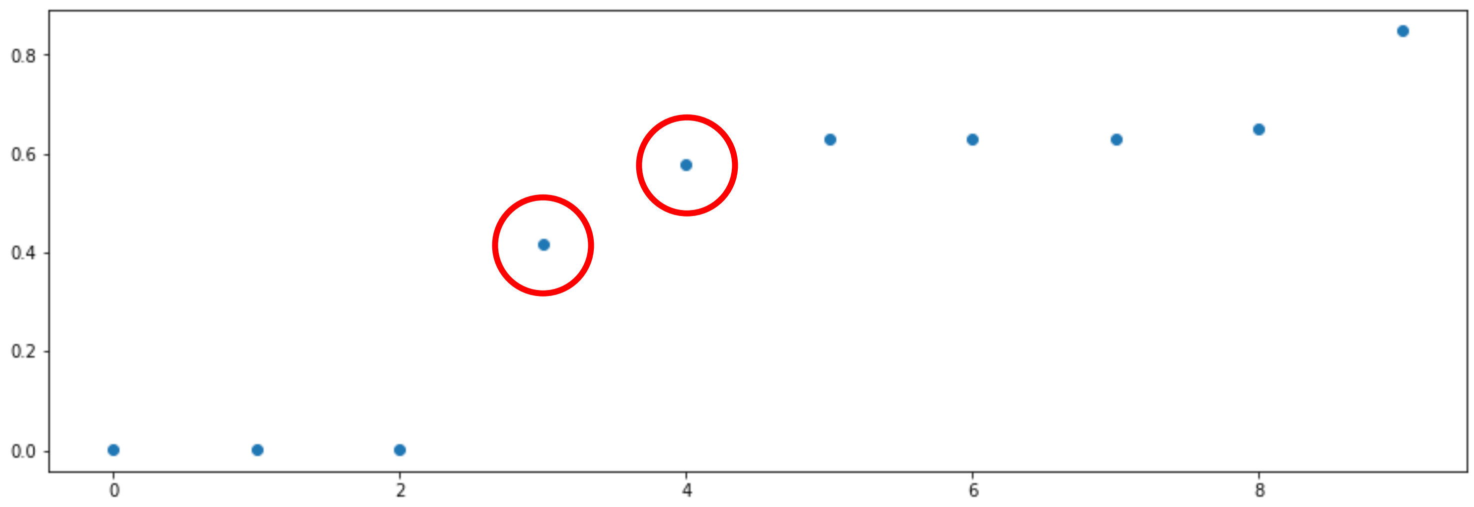
→ Perform large scale clustering using KMeans

→ Examine clustering performance using elbow method

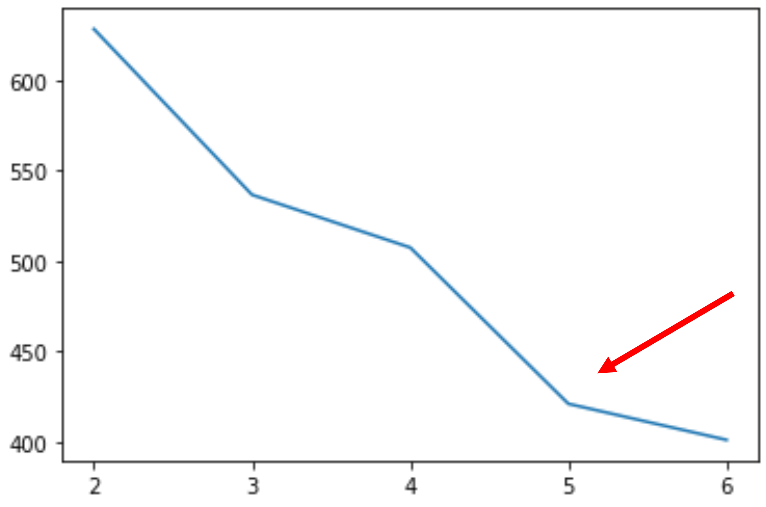
Customer Segmentation

Objective	▪ Use unsupervised learning to segment customers into groups

Largest increases in eigenvalues



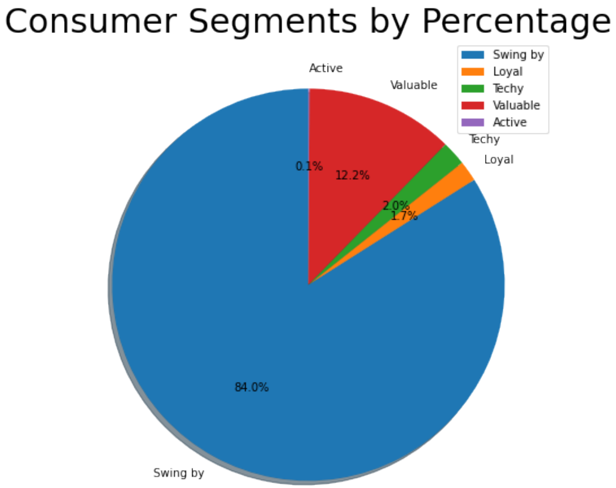
Minimized inner cluster distance

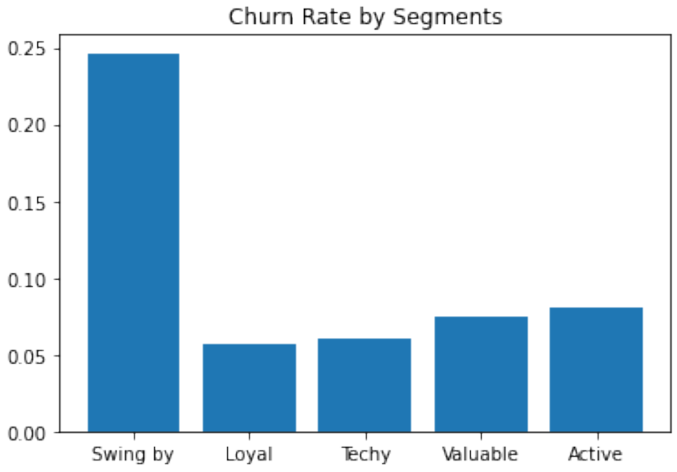


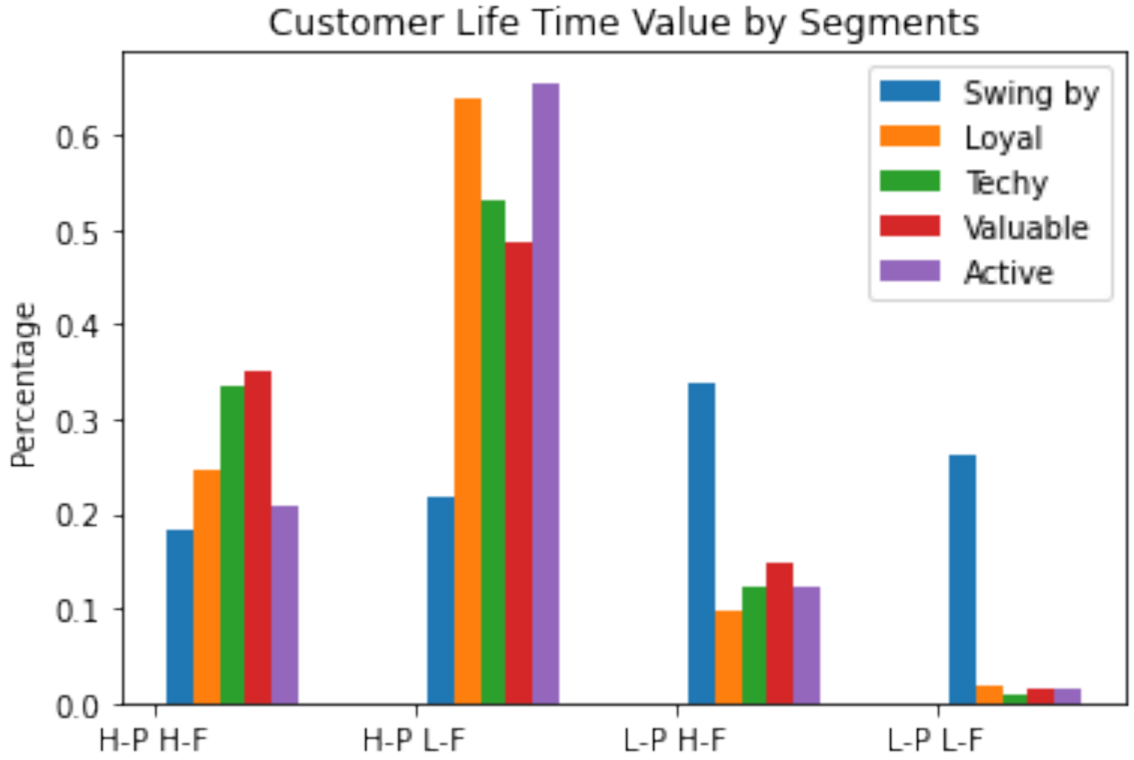
Customer Segmentation

Segment	Characteristics
“Swing by”	**Highest churn** | Lower LTV | Lowest average mobile logins | Use Venmo/PayPal the least | Highest % of closed accounts | Lowest number of remote deposits
“Loyal”	**Lowest churn** | Highest number of calls to call center | Highest average age in households | Highest number of saving accounts | Highest % of high income
“Techy”	Highest average mobile logins | Highest % Uber/Lyft payments | Use Venmo/PayPal the most
“Valuable”	Largest percentage of H-P H-F | Highest direct deposit amounts | Lowest amounts of check deposits | Fulton customer the shortest | Younger households (often Gen X) | Highest % of middle income
“Active”	Highest average of billpay transactions | Highest average (deposit, investment, loan) products in household

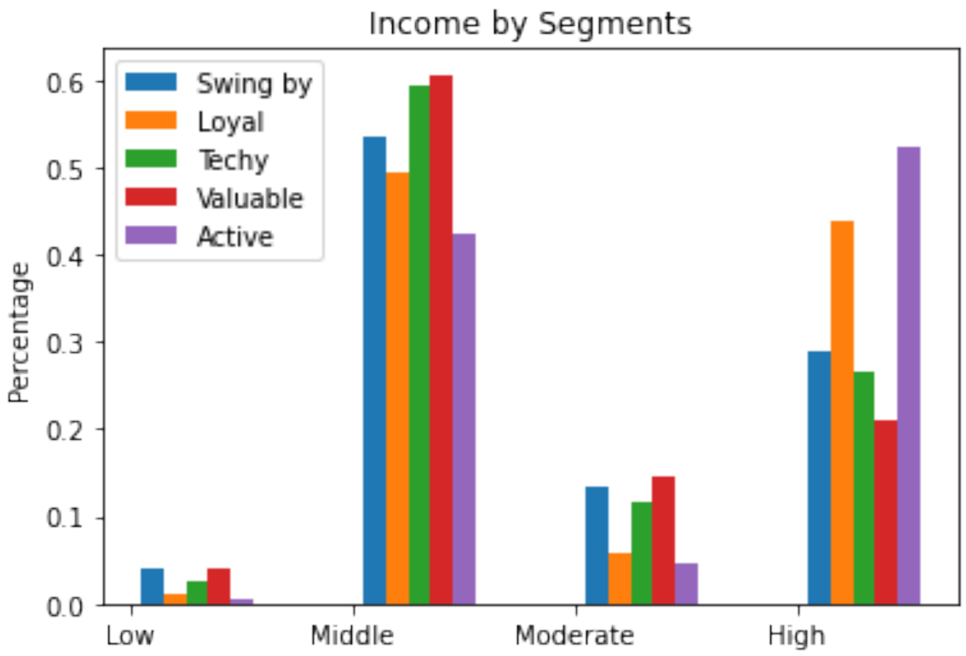
Customer Segmentation

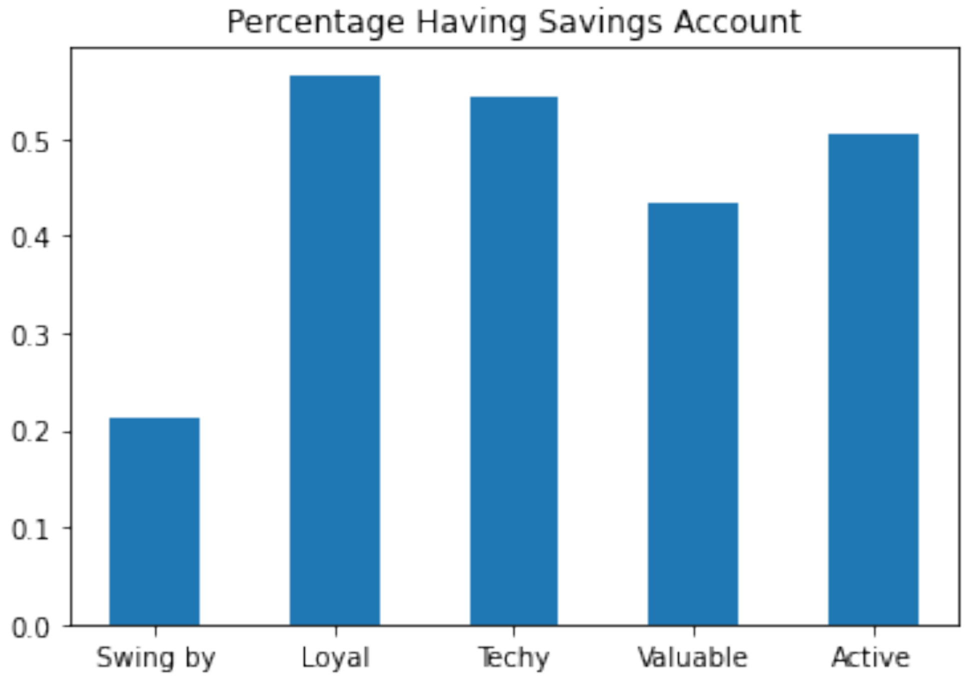




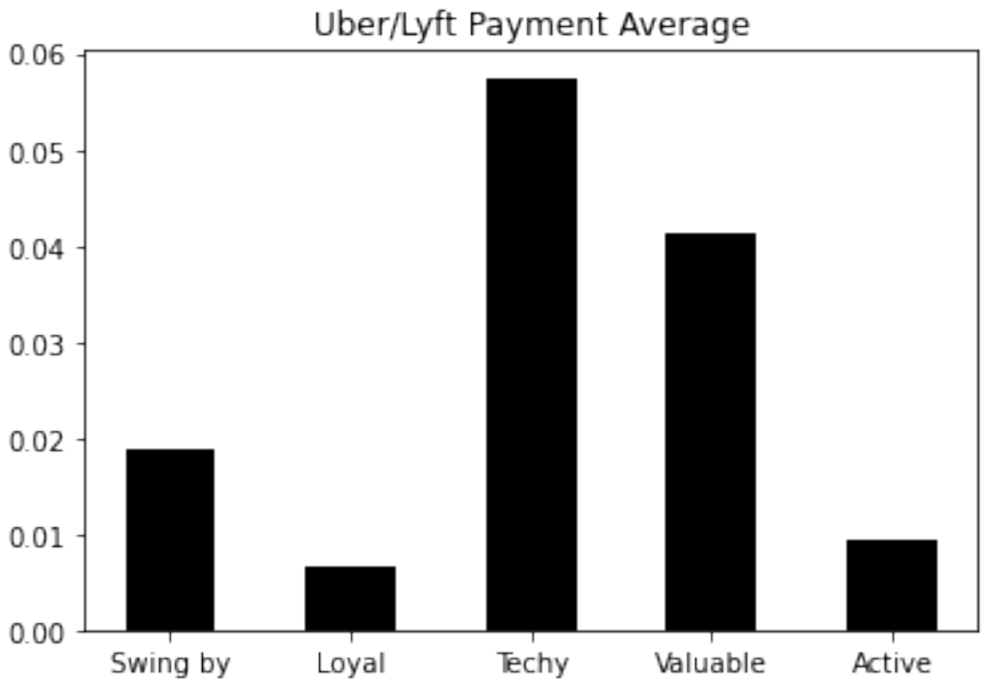


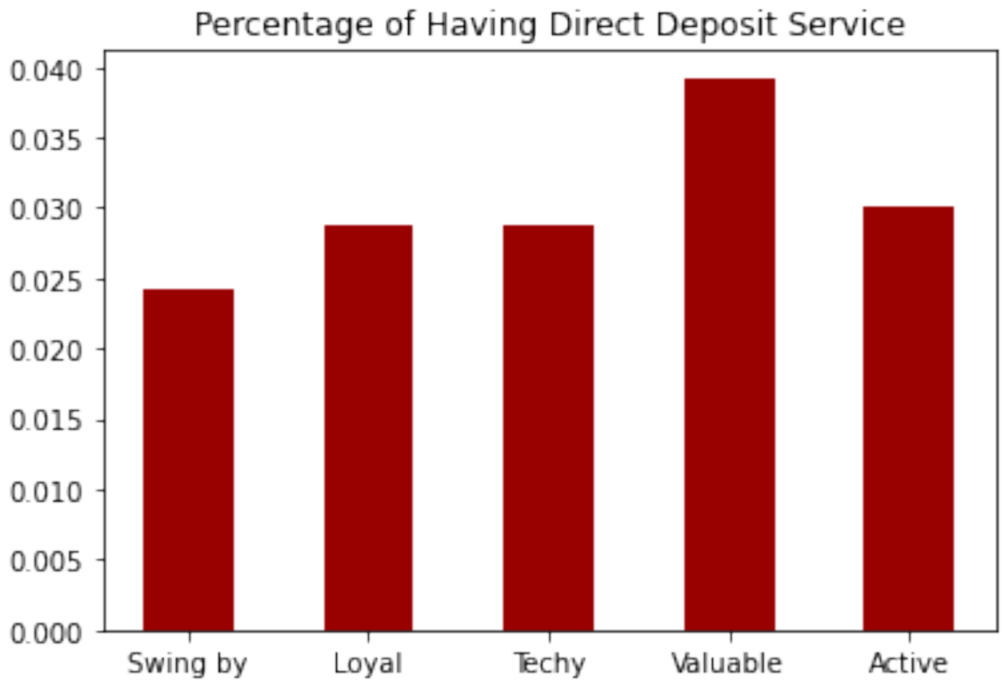
Customer Segmentation





Customer Segmentation





Business Recommendation: Target Customer Segments

Objective

- Classify customers into segments to provide customers with targeted recommendations that meet their needs and increase loyalty

Acquisition

- Identify and execute campaigns targeting customers with characteristics of low-churn segments

Servicing

- Provide services tailored to the customer's needs based on segment traits
- Recommend or cross-sell products associated with loyalty

Relationship

- Build in-depth relationships with customers via analytics-based personalized services

Retention

- Prioritize converting "swing by" customers to other segments with lower churn

Feature Scoring Procedure

Objective

- Which features are giving the most improvements to accuracy in a nonlinear model?

125 Choose 2 = 7750 Columns

Fit 7750 Random Forests
(Each column will be in 124 of the models)

Distribution of average accuracies for each model that a column participated in

Scored column as the equally weighted average of the mean, median, 90th percentile, and max of the distribution

Graphed column scores and picked a natural cutoff point

Feature/Column Scoring Results

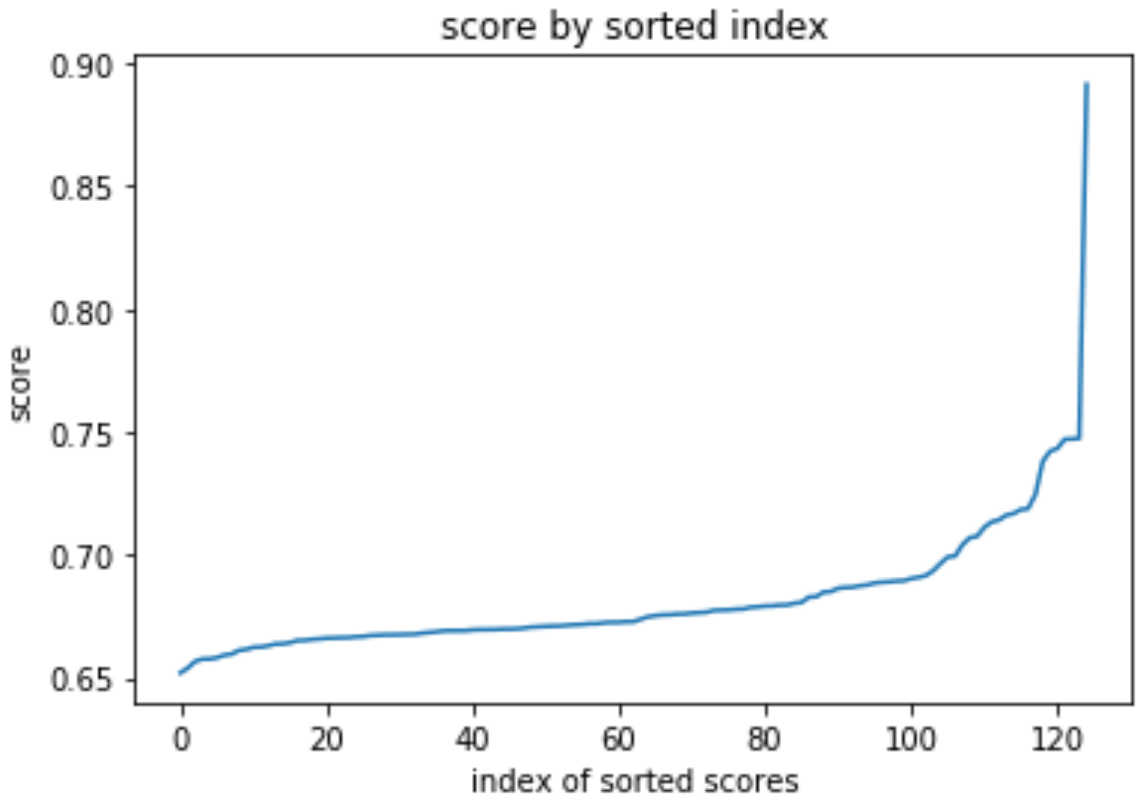
Top Scoring Columns

Index	name	score	mean	max	90%ile	median
6	STARTPROD	0.891464	0.876033	0.909908	0.899672	0.880242
63	TOTAL_ASSETS_FFC	0.747434	0.693626	0.909908	0.69857	0.687631
82	hascheckingactivity	0.7472	0.701238	0.878022	0.714385	0.695153
7	NEWPROD	0.747086	0.702596	0.856132	0.73875	0.690867
43	BROKERAGEBAL	0.743331	0.692753	0.895055	0.698707	0.686809

Bottom Scoring Columns

Index	name	score	mean	max	90%ile	median
102	tot_calls	0.652	0.571116	0.821033	0.649147	0.566702
10	MOVEDHH	0.653974	0.566885	0.836254	0.648787	0.563971
53	DEPO_SRV_TOT	0.656639	0.620682	0.74392	0.654359	0.607594
15	IRACONSRV	0.657627	0.571767	0.841784	0.649831	0.567124
16	BROKERAGESRV	0.657671	0.567384	0.850581	0.648758	0.563961

Score by Sorted Column Index



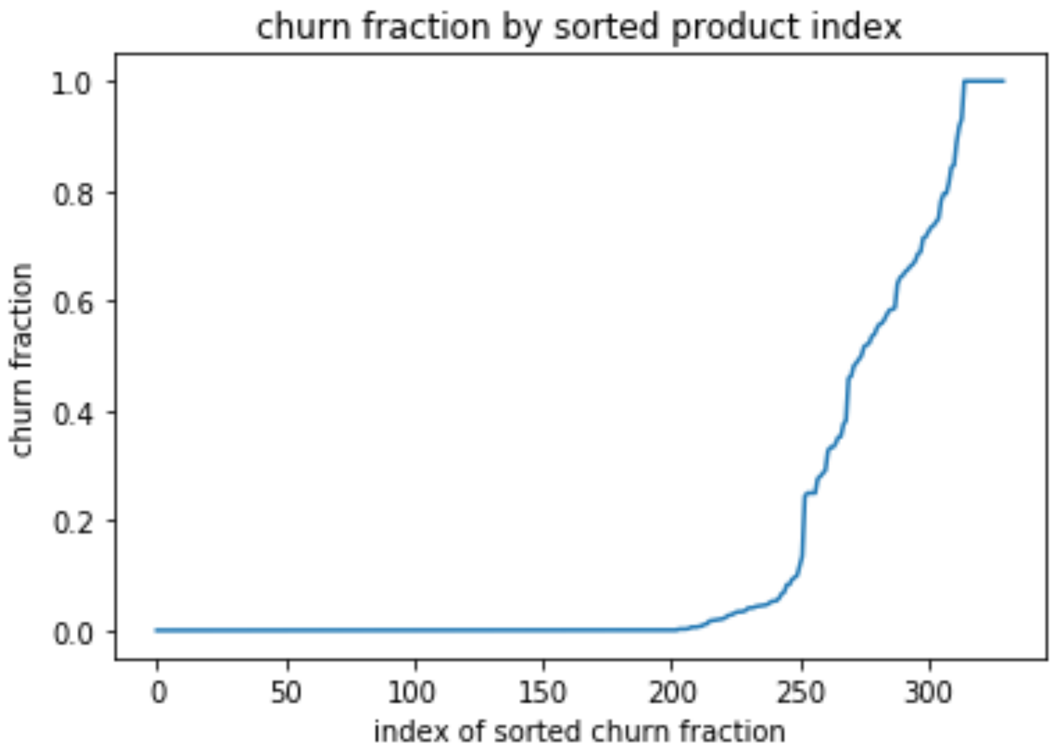
Starting Product

Most vs. Least Likely to Churn

STARTPROD	% churn	count
TTAC	1	99
TUNA	1	12
IRAF	0.931148	305
OLB	0.916667	12
CDPB	0.888092	2091
MMPER	0.846154	2964
FTAC	0.843844	333
LCIND	0.814502	12895
MTGS	0.79519	7734
EQOPT	0.795181	581
EQMTG	0.782609	92
BUNP	0.75	12
CKINT	0.74317	22073
TB	0.736864	2398

STARTPROD	% churn	count
FILN	0.00740398	2161
NLCN	0.00598802	167
A	0.0059761	502
HILN	0.00591716	169
FMLN	0.00371747	269
AILN	0.00255754	391
FVCCATM	0.00240096	833
VD	0.00236967	2110
NILN	0.00220751	453
ND	0.000547645	5478
DD	0.000161838	6179
GD	0.000116414	25770
AATMATM	0	46
ACLNA	0	14

Churn Fraction by Product



Predictive Model Implementation

Objective	▪ Design a model that predicts whether a household churns or is kept

Method Implementation Notes

- Nonlinear model allows for complex interaction
- By using “class_weight = ‘balanced’” in the model, we make sure the accuracy on the kept households and churned households are prioritized equally
- Since there are fewer churned HH, the precision suffers, but this is in line with business intuition of losing a customer is more expensive than the cost to keep an existing customer from churning

Churn Model Metrics

	Accuracy-Kept HH	Accuracy-Churned HH	Precision	Recall	F-score
Random Forest (sqrt)	0.867	0.94	0.656	0.94	0.773
Logistic Regression w/ L2	0.408	0.848	0.28	0.848	0.421
Random Forest (all)	0.792	0.991	0.564	0.991	0.719
F.S. Random Forest (sqrt)	0.866	0.933	0.655	0.933	0.769
F.S. Random Forest (all)	0.793	0.991	0.565	0.991	0.72
AdaBoost	0.988	0.92	0.954	0.92	0.936
F.S. Adaboost	0.983	0.922	0.935	0.922	0.928

- Random forest “sqrt” vs “all” refers to checking sqrt(features) or all features at each split
- F.S means the model is run on feature selected data - the top 35 rows
- Logistic regression has poor performance, but we may be able to get more meaningful significance data out of it.

Business Recommendation: Utilize Predictive Variables

Objective

- Use variables most predictive of churn to inform insights and strategies personalized for the customer

Understand

- Examine intuition for starting product and other high-scoring variables
- Improve data tracking to include more of customers product profile

Predict

- Ensure data is robust enough to draw conclusions
- Use a nonlinear model to predict whether customers are likely to churn

Strategize

- Reconsider profitability of high-churn products
- Encourage customers to switch to or add low-churn products

Conclusion

Goals	
	▪ Use the consumer dataset ▪ Segment the Fulton Bank customer base ▪ Build a model that predicts customer churn

Recommendations

- **Customer segmentation**
  - Use customer characteristics to segment customers, allowing for easier acquisition, servicing, relationship development, and retention of customers
- **Predictive model**
  - Predict the likelihood of churn in an individual customer
  - Formulate strategies based on a trait’s association with high or low churn