Hotel Customer Segmentation

File created on: 9/9/2022 from Tableau Public

Lindsey Robertson

Customer Segmentaton & Attrition Prediction

September 2022

Customer Retention

* Existing customers spend more and cost less
* Focusing on exiting customers fosters loyalty and free advertising

Revenue Opportunity

1. How can we understand the customer and their value better?
2. Can we predict when a customer is at risk of never retunrning in order to appropriately intervene?

*click anywhere to visit GitHub Repo with all notebook processes and more EDA visualizations

Current State

Attrition	
False	True
22,410,469	8,216,173
44,230	19,440

Total Customers (with confirmed booking)
63,670

Total Revenue
30,626,642

Assumed Avg. Annual Revenue 10,208,880

Attrition Rate Over 3 Years
44%

Assumed Avg. Annual Attrition
15%

Data:

3 years of aggregated customer booking data with demograp, segment and behavioral information querried from a SQL Database. Stored in flat file from Kaggle here: https://www.kaggle.com/datasets/nantonio/a-hotels-customers-dataset

* Over 80K Observations
* 31 Features

Original Features

F3	F6	F2
ID	int64	0.00
Nationality	object	1.00
Age	float64	2.00
DaysSinceCreation	int64	3.00
NameHash	object	4.00
DocIDHash	object	5.00
AverageLeadTime	int64	6.00
LodgingRevenue	float64	7.00
OtherRevenue	float64	8.00
BookingsCanceled	int64	9.00
BookingsNoShowed	int64	10.00
BookingsCheckedIn	int64	11.00
PersonsNights	int64	12.00
RoomNights	int64	13.00
DaysSinceLastStay	int64	14.00
DaysSinceFirstStay	int64	15.00
DistributionChannel	object	16.00
MarketSegment	object	17.00
SRHighFloor	int64	18.00
SRLowFloor	int64	19.00
SRAccessibleRoom	int64	20.00
SRMediumFloor	int64	21.00
SRBathtub	int64	22.00
SRShower	int64	23.00
SRCrib	int64	24.00
SRKingSizeBed	int64	25.00
SRTwinBed	int64	26.00
SRNearElevator	int64	27.00
SRAwayFromElevator	int64	28.00
SRNoAlcoholInMiniBar	int64	29.00

Created Features

* CTRY - top 15 "Nationalities" category else "Other"

* Total Reveune - Lodging Revenue + Other Revenue

* RFM Segments - Quanitle Calculated

* RFM KMeans Clusters

* Person Kmeans Clusters

* Attrition T/F - "DaysSinceLastStay" threshold greater than 2 years

Assumptions

* Profiles are created for each guest companion(adult or children).

*Customer profile is created by one of three things:
- customer's first checked-out at the hotel
- customer's first cancelation
- customer's first no-show

* Sometimes there are more than one profile for the same customer.

* Only after the customer's first stay can hotels confirm the guest's personal details, such as nationality.

Data Cleaning Resolutions:

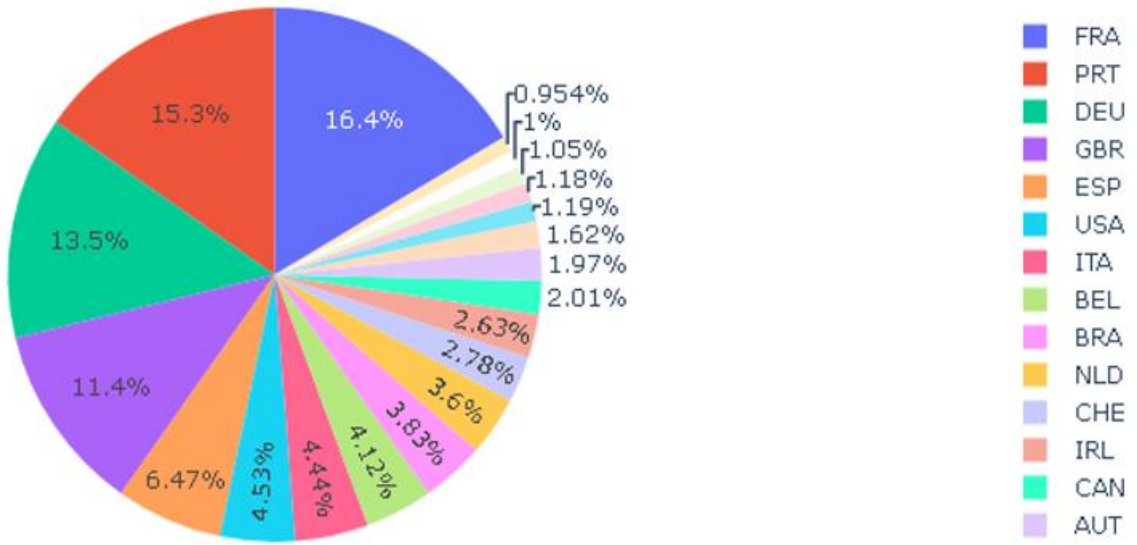
*** Over 3k 'Age' values left with 0 value, assuming blank or error input**

*** 25 'Age' values were negative value or over 100. Imputed to 0.**

*** Over 20k instances with no monetary value were removed as they can be considered in modeling when they become a paying customer. New customer aquisition of this cohort would be a separate project.**

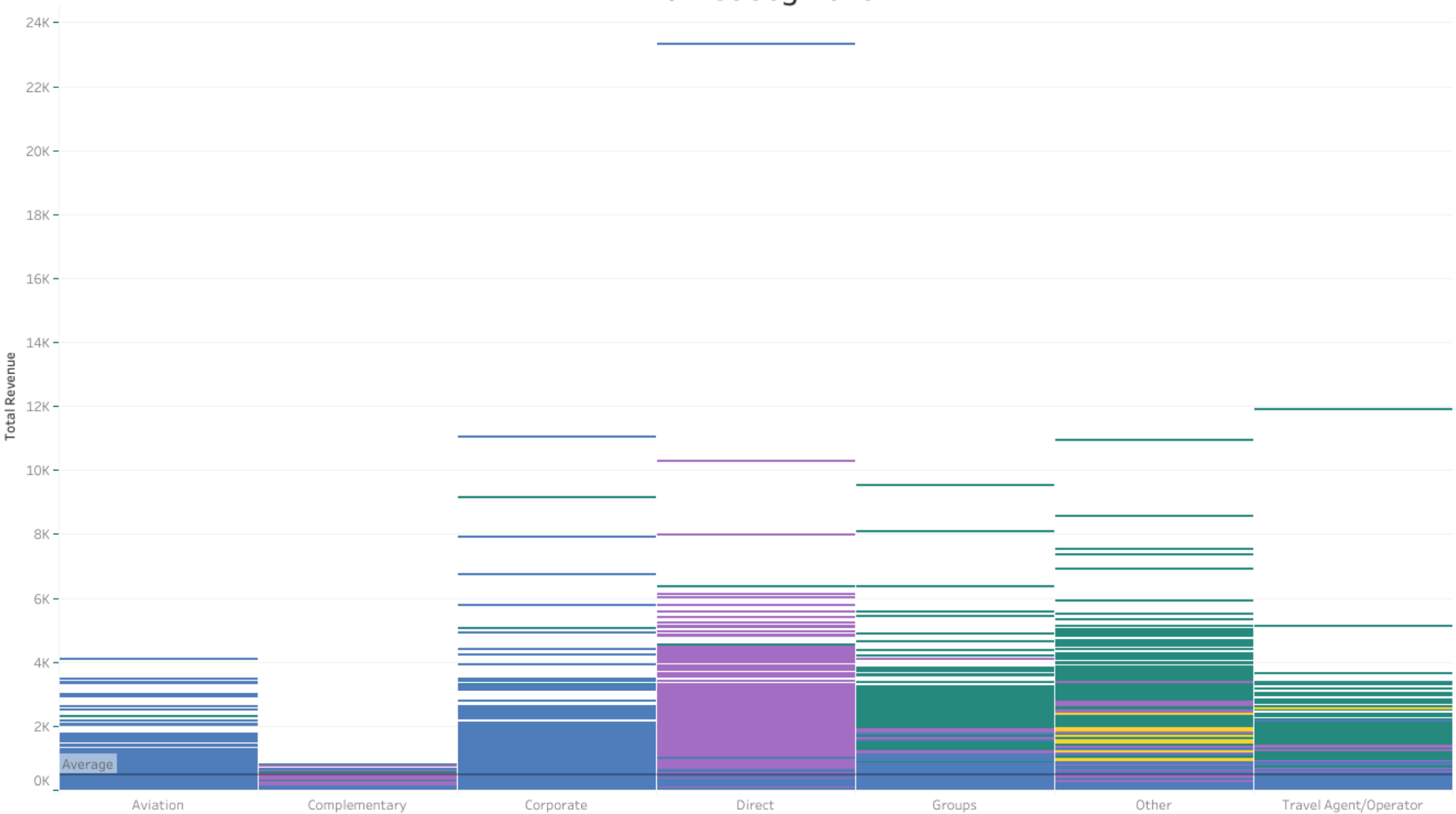
*** Removed 'Hash' features.**

*** Consolidated 'Nationalities' into top 15 Countires in frequency (over 1000 customer records) else 'Other'.**




Reveuene from top Countries with the most guests
Total Revenue
423,097
5,393,980
Russia
Mongolia
China
Japan
Indonesia
Greenland
Mexico
Venezuela
Colombia
Peru
Norway
Finland
Poland
Belarus
Ukraine
Turkey
Iraq
Iran
Algeria
Libya
Egypt
Saudi Arabia
Sudan
Yemen
Democratic Republic of Congo
© 2022 Mapbox © OpenStreetMap

Market Segment

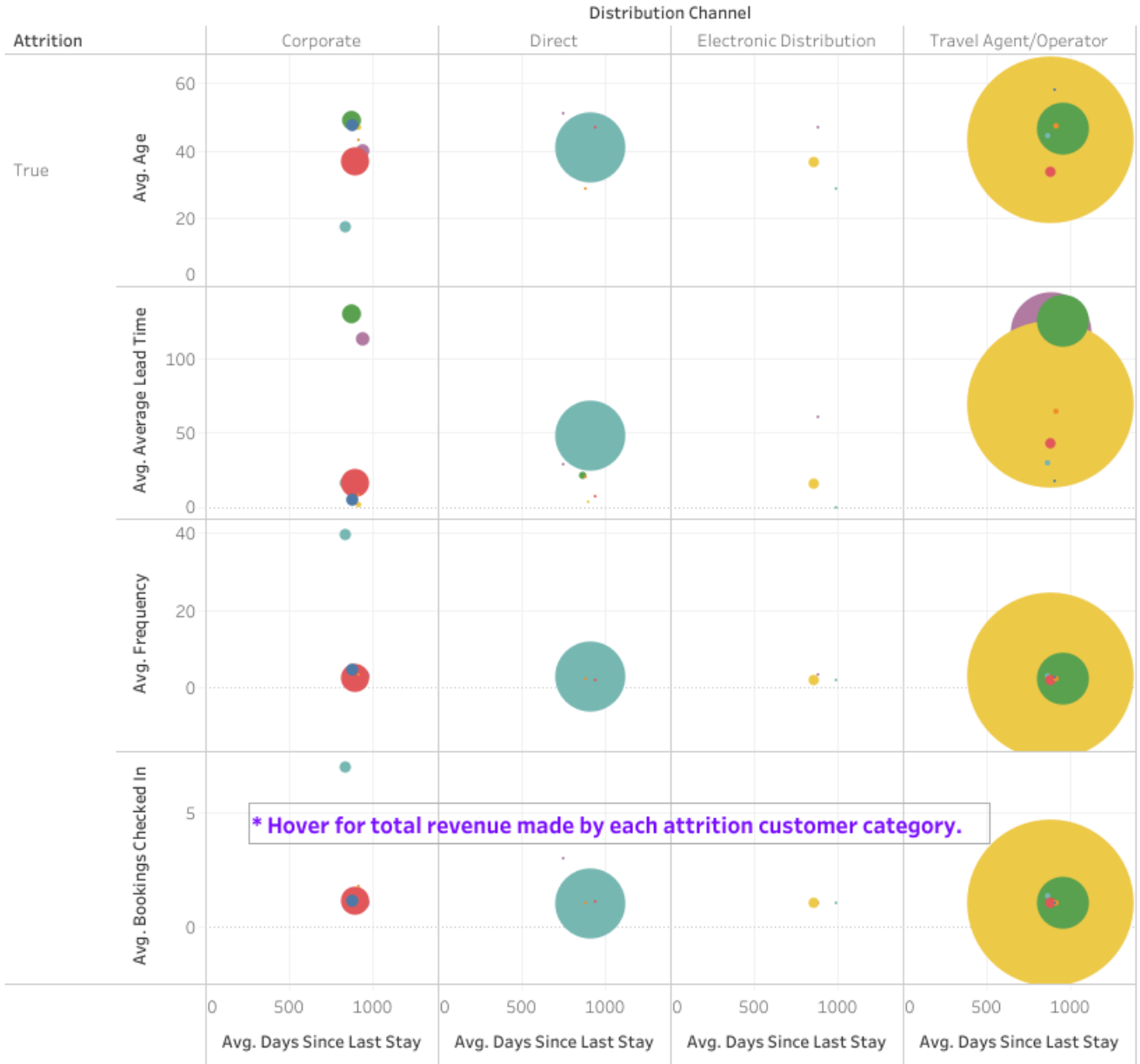


Distribution Channel
- Corporate
- Direct
- Electronic Distribution
- Travel Agent/Operator

Highlight Distribution Ch..
No items highlighted

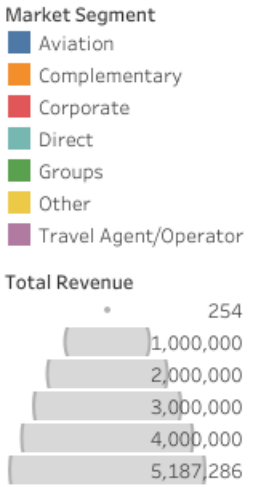
Highlight Market Segment
No items highlighted

***Hover for 'Age' and 'Country' information**



Days Since Last Stay Correlations

Filter for only customers labeled with 'True' for Atrrition



Data Considerations:

* Roughly 30% of the dataset are new customers with little history

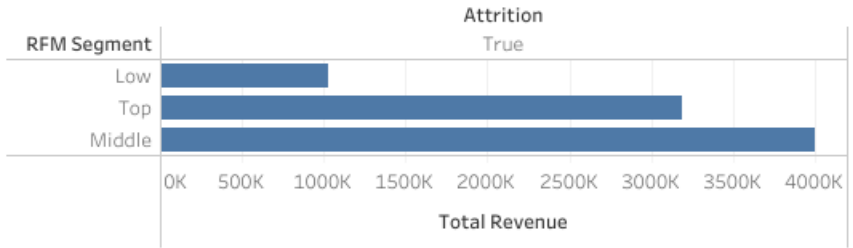
* Most key features are skewed

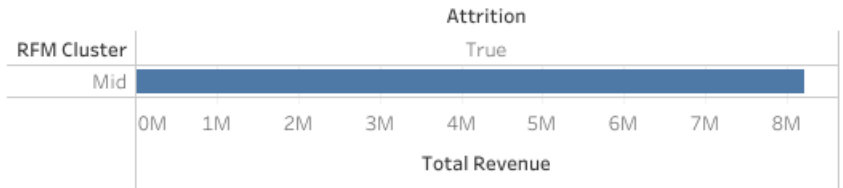
* Heavy imbalance in many key categories

* Almost 20k customers over 730 days (lost customer of over 2 years)

* If just 5% were viable intervention targets, we could save roughly 950 customers at their various average revenue.

* 5% of total past attrition reveues would be $400,000 in hypothetical missed revenue over those 2 years





RFM Analysis of Lost Customers

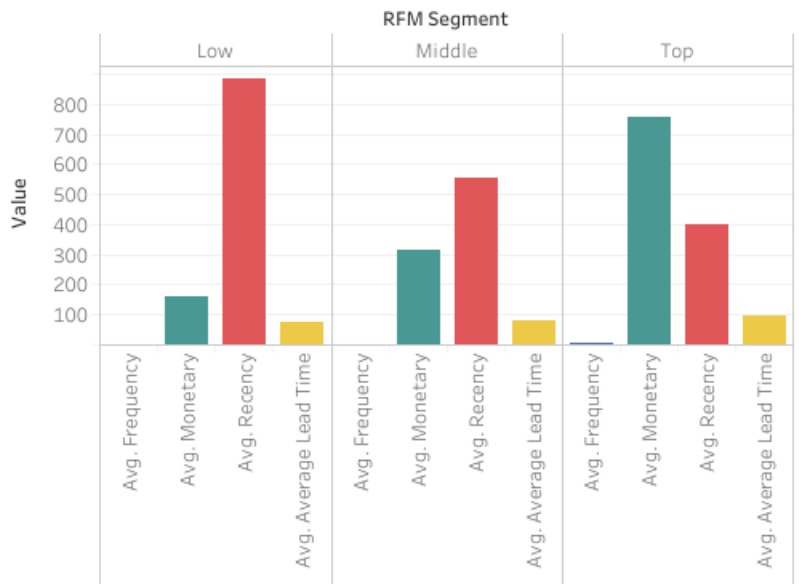
Recency = how long since the customer stayed

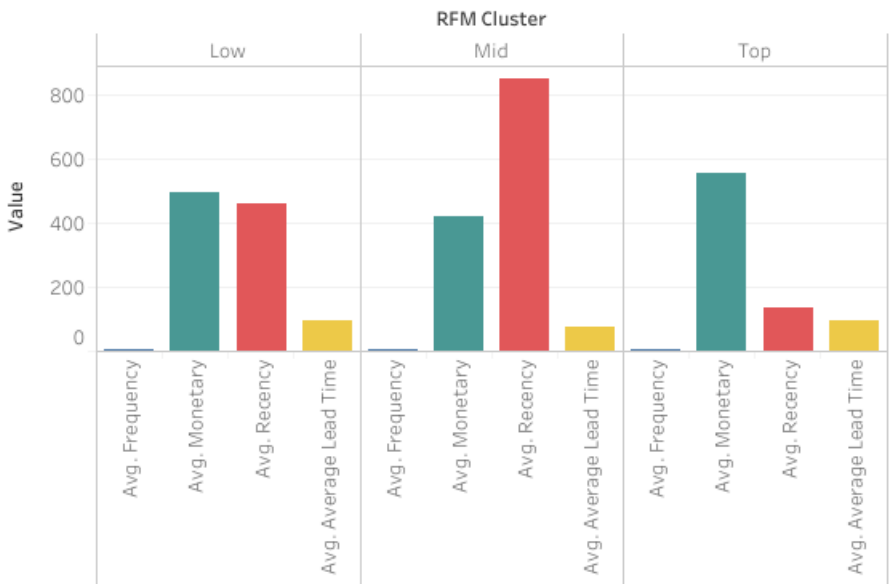
Frequency = how often does the customer stay based on RoomNights (#of rooms X # of nights)

Monetary = total reveue of customer

Quanite Segments - Create an RFM scoring metric to segment customers by value quantiles

Clustered Segments - Kmeans Clustering



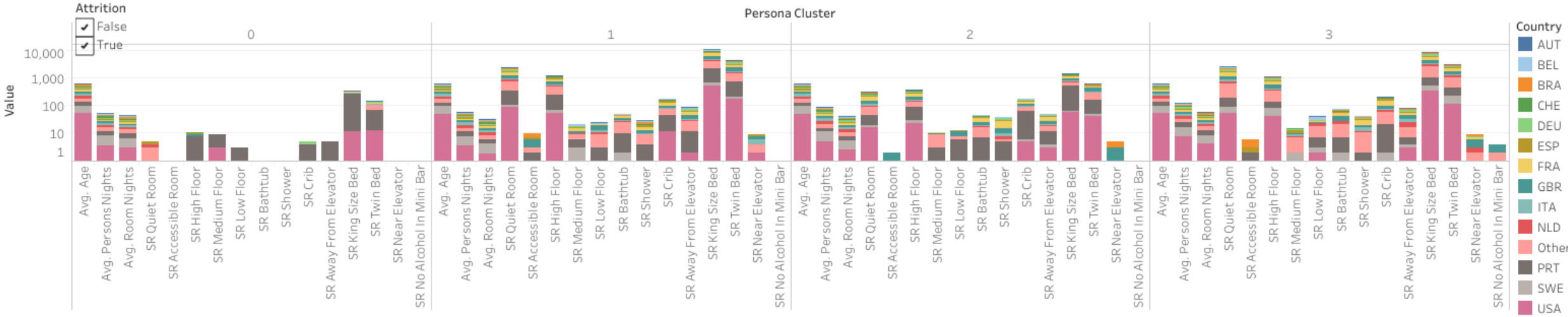


Considerations:

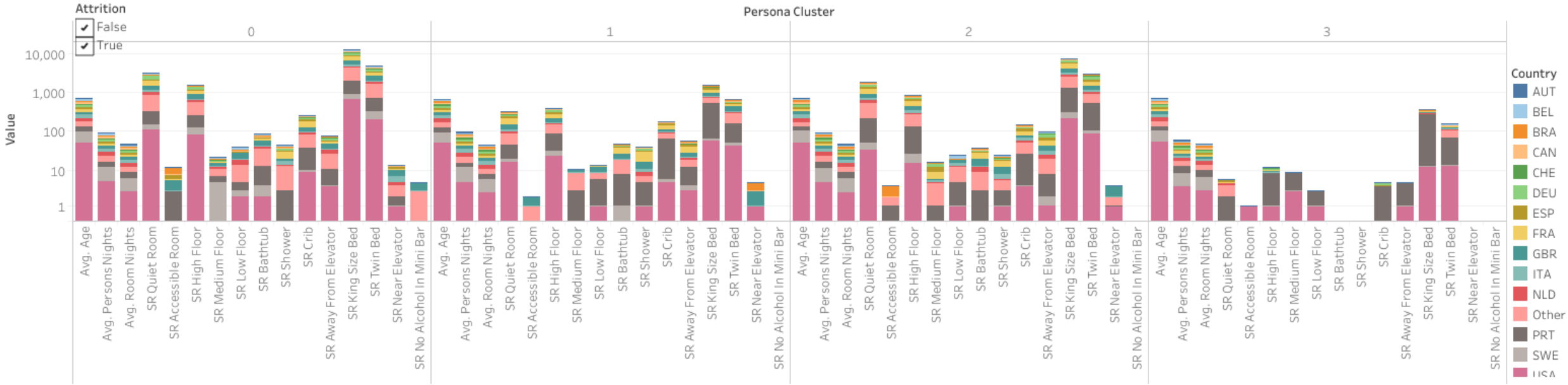
*20k records with recency value -1. Imputed to zero.
* Removed 20k records from model and anlysis without revenue data
* Log Transform, Cener and Scale RFM values only for better visualization and clustering ability.

KMeans Cluster Analysis of Customer Behavior

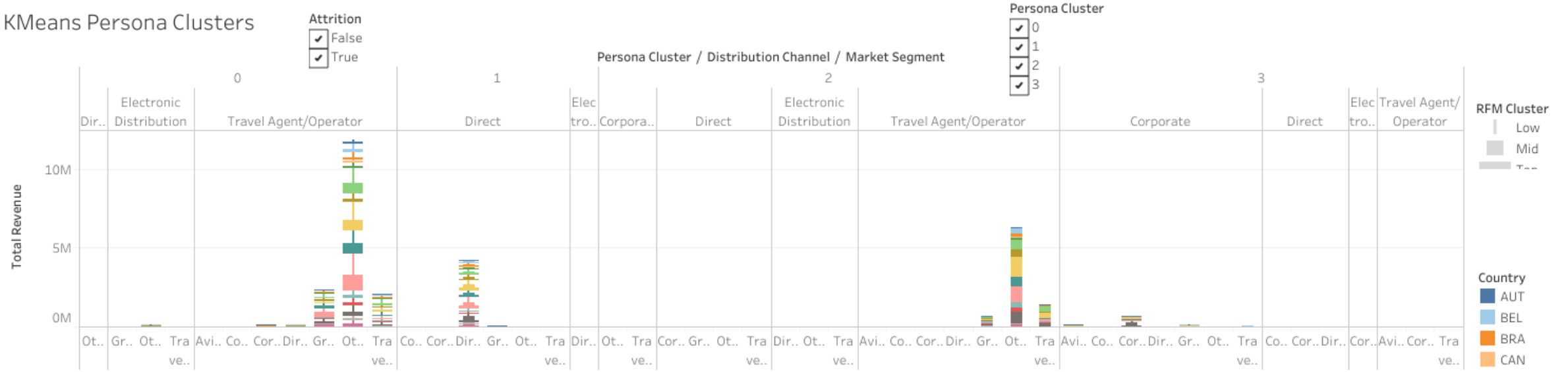
Persona Clusters



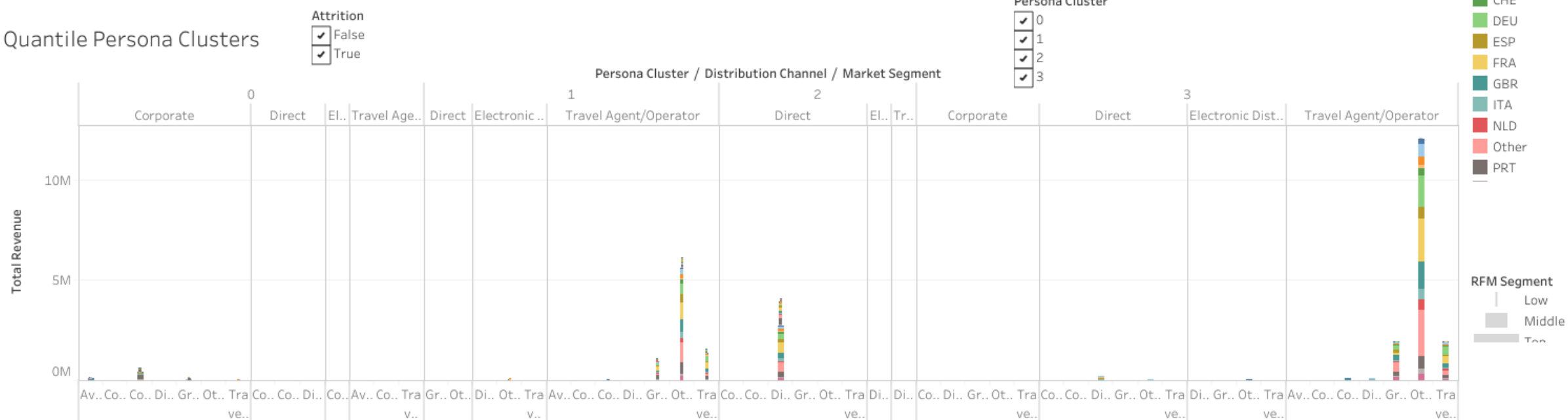
Persona Clusters Quant



KMeans Persona Clusters



Quantile Persona Clusters



KPIs

* Reduced 'DaysSinceLastStay'/'Recency'
* Reduced Attrition Rate
* Increased Bookings
* Increased Revenues

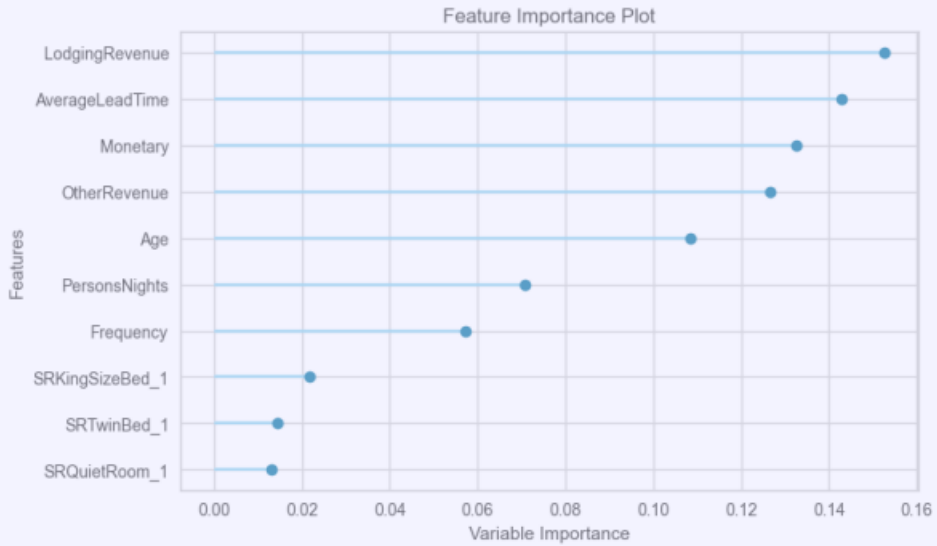
Top Metrics

* Recall to limit false negatives
* AUC to measure ability to predict classes correctly overal. Best model fit metric.

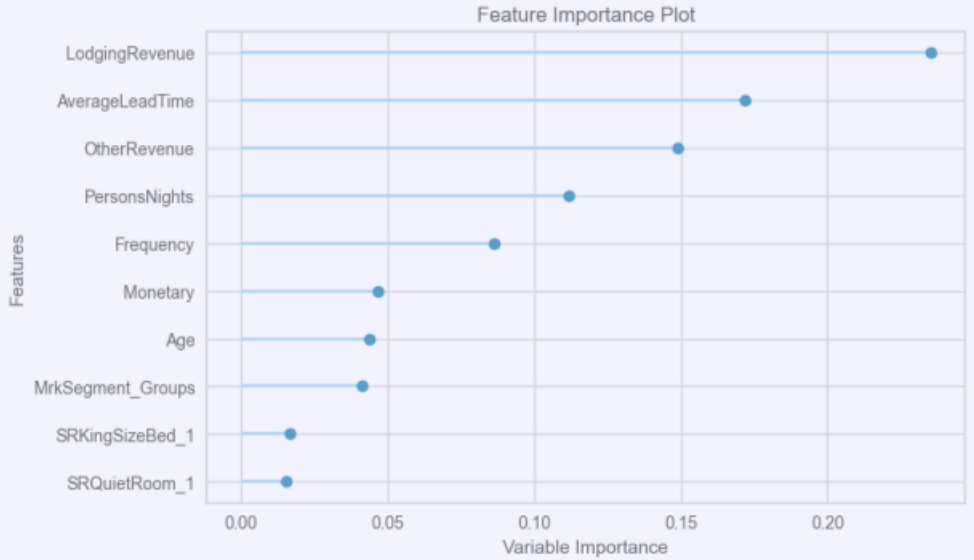
Final Experimentation Metric Comparison

		AUC	Recall	Accuracy	Precision	F1
Mean	tune_py_OptimizedRF	0.8385	0.6615	0.7723	0.6291	0.6448
	stacked_best	0.8385	0.6615	0.7723	0.6291	0.6448
	Baisc Random Forest	0.8385	0.6615	0.7723	0.6291	0.6448
	boosted_best	0.8356	0.7743	0.7361	0.5558	0.6470
	tuned_blended_best	0.7816	0.6266	0.7204	0.5459	0.5834
Std	tuned_blended_best	0.0065	0.0084	0.0069	0.0106	0.0072
	tune_py_OptimizedRF	0.0053	0.0078	0.0062	0.0117	0.0073
	stacked_be	0.0053	0.0078	0.0062	0.0117	0.0073
	Baisc Random Forest	0.0053	0.0078	0.0062	0.0117	0.0073
	boosted_best	0.0037	0.0100	0.0054	0.0071	0.0054

Base Estimator - RF



Threshold Optimized Estimator- RF



Best Approach

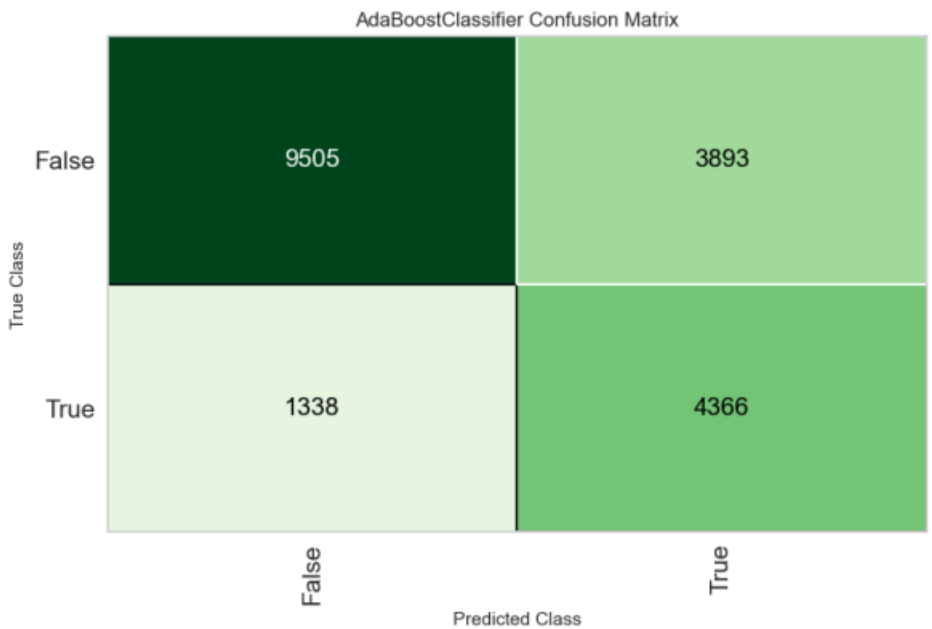
1. Remove all features related to RFM in any way to avoid leakage
2. Remove persona clustering features to prevent leakage
3. Remove highly correlated features to target variable
4. Encode categorical features with dummy features
5. Confirm more multicollinearity adjustments are not needed.

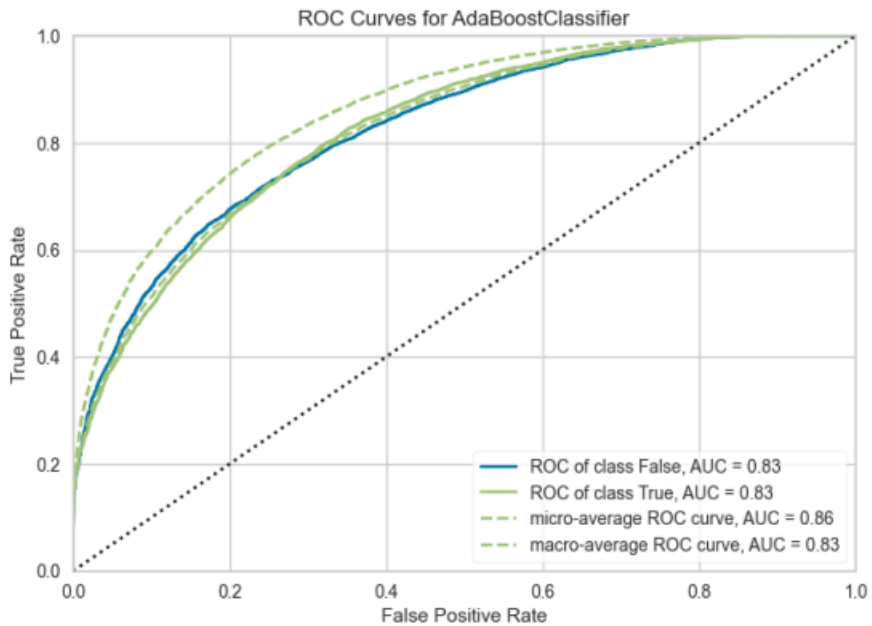
Best Model

AdaBoost algorithm to focus on errors, applied to Random Forest base estimator with custom probability threshold of .40 to favor Recall.

Parameters

* Implement overesampling using SMOTE for slight imbalance in the target feature
* Normalize the data distribution using z-score on non-encoded features
* Remove outliers witiin 2.5% of either side.
* Cross validate on training data for evaluation before test predictions.
* 'Boosted' ensembling method applied to the Random Forest base estimator
* Probability threshold of .40, optimized for Recall, applied to create: meta_estimators.CustomProbabilityThresholdClassifier.
* 10 n_estimators trained sequentially.
* The boosting ensemble helped the model focus on incorrect predictions.





Model Implications:

*If only 10% of correctly identified customers at risk of attrition were retained at the average revenue, intervention could create roughly 185,000 in potential revenue over the next 3 years.

* This figure can be sharpened by calculating average revenues for the different segments. (See RFM Segementation and Persona Cluster Segement slides for revenue dashboards.

* Data snapshots will allow us to adjust these figures for specific performance periods.

* Other recomendationscan create better customer loyalty and experience, which will have even more positive impacts on revenue growth.

Recommendations:

1. Add data points for more robust features:
   - daily, weekly, monthly, annual snapshots with timestamps for more accurate CLV performance periods, seasonality analysis and for measuring the KPIs.
   - add a feature to note if the customer paid upfront for future cancellation or noshow prediction.
   - add propensity score to the attrition prediction for levels of targeted action.

2. Monitor KPIS more precisley with new features on dashboards
   - Attrition Rate
   - Churn/Cancel/NoShow Rate
   - New Bookings Rate
   - Revenue Change

3. Work on identifying duplicate profiles with address field feature

4. Create more user-friendly customer behavior dashboards for marketing and managemnt to develop their targeted action plans.

5. Connect any satisfaction survey data as a key feature of attrition.

5. Develop cost of intervention and aquisition fileds based on segment and persona cohorts.

6. Explote a loyalty program cost/benefit.

Use Case

* Segment analysis allows new features to be derived that will inform more targeted promotions and marketing.

*Attrition predictions can allow for propensity scores and additional features that can be flagged for managment and marketing to act on for customer retention, including behavoior insights about segments.

* Dashboards implemented for monitoring sucecss of the project by better measuring business KPIs.

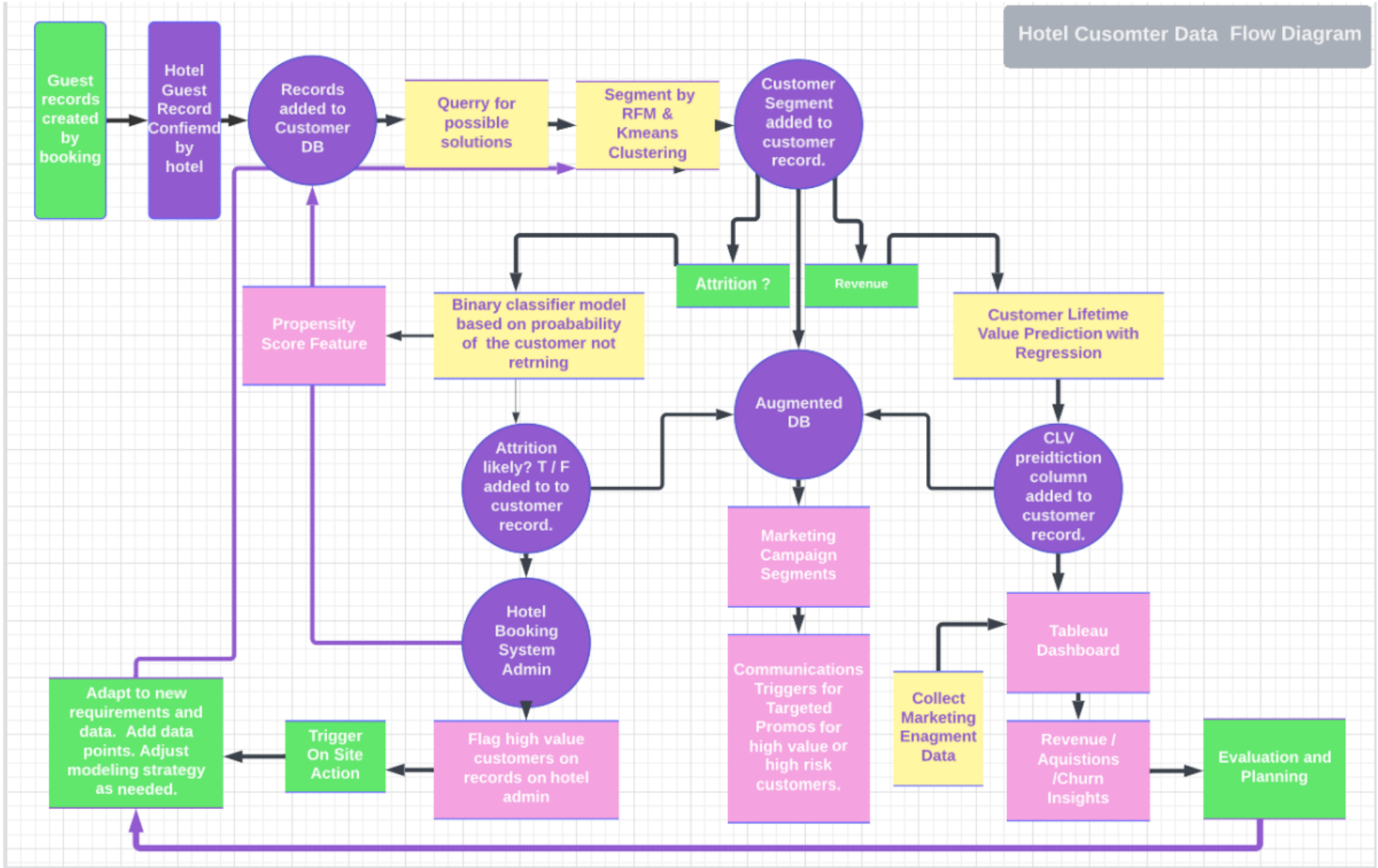
User Case

* **Reception & Hotel Mangement**- high value customers flagged for hotel staff to provide upsale opportunities and priority service.

***Marketing** - dashboards can inform campaign and promotion development from a cost and customer preference perspective.

***Trigger altert for intervention** campaign to customers who are at risk of never returning.

* **Executives** - dashboards that inform their planning and spending effort and measure the success of the solution to adapt.


Hotel Cusomter Data Flow Diagram
Guest records created by booking
Hotel Guest Record Confiemd by hotel
Records added to Customer DB
Querry for possible solutions
Segment by RFM & Kmeans Clustering
Customer Segment added to customer record.
Attrition ?
Revenue
Binary classifier model based on proabability of the customer not retrning
Propensity Score Feature
Customer Lifetime Value Prediction with Regression
Augmented DB
Attrition likely? T / F added to to customer record.
CLV preidtiction column added to customer record.
Marketing Campaign Segments
Hotel Booking System Admin
Tableau Dashboard
Communications Triggers for Targeted Promos for high value or high risk customers.
Collect Marketing Enagment Data
Adapt to new requirements and data. Add data points. Adjust modeling strategy as needed.
Trigger On Site Action
Flag high value customers on records on hotel admin
Revenue / Aquistions /Churn Insights
Evaluation and Planning