RISING ANALYTICS
Data. Insight. Action.

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Donor Score and Optimal Ask

Use Case: Clark Atlanta University (CAU)

Rising Analytics worked closely with the CAU Alumni Relations Department. The CAU Alumni Relations Department works to establish, build, maintain, and strengthen the relationships with Alumni constituents, and drive alumni donations to the university. The team is also responsible for recording and maintaining alumni information, coordinating with the national alumni association and affiliates, and facilitating requests associated with alumni communications, events and volunteers.

Table of contents

RISING ANALYTICS
Data. Insight. Action.

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Problem Statement

We worked closely with the CAU Alumni Relations Department whose primary job is to identify the best donor from the alumni pool, identify different marketing strategy to approach different alumni, ask for optimal donation amount and so on. They had around 125K alumni records. Through this study, we hoped to develop some insights that can help organization to hold a marketing strategy in place that can work differently for different segment of alumni.

Objective and Scope of the project

1. Objective
   The primary objectives of the study are:
   - Classify Donors and Non-donors
   - Providing score/probability to all the alumni. Better the score more the probability of alumni becoming donor.
   - Providing alumni segmentation based on the behavior.
2. Scope
   - The scope of the study covers 130K Alumni records.
   - The study covers 100 Years of data starting from Jan-1920 to Dec-2019
   - The study focuses only on the demographic variables provided by client.

Data Source

The data were collected from the Blackbaud raiser's edge platform.

Tools and Techniques

We have used following analytical techniques / methodology for analyzing data:

1. Summary of Statistics for each variable
2. EDA. Using Graphs and plots to visually represent all the variables.
3. Identification of significant variables through correlation matrix and WoE/IV.
4. Apply statistical as well as machine algorithms for alumni classification.
5. Apply statistical and time series forecasting for optimal ask.
6. Tools Used: R, Python, MS Excel, PostgreSQL, and Power BI
7. Techniques: Box Plot, Histogram, Bar Chart, Line Chart, Correlation Matrix, Logistic Regression, Random Forest, GBM, Linear regression, RFM segmentation analysis, Time series forecasting

Analytics Approach

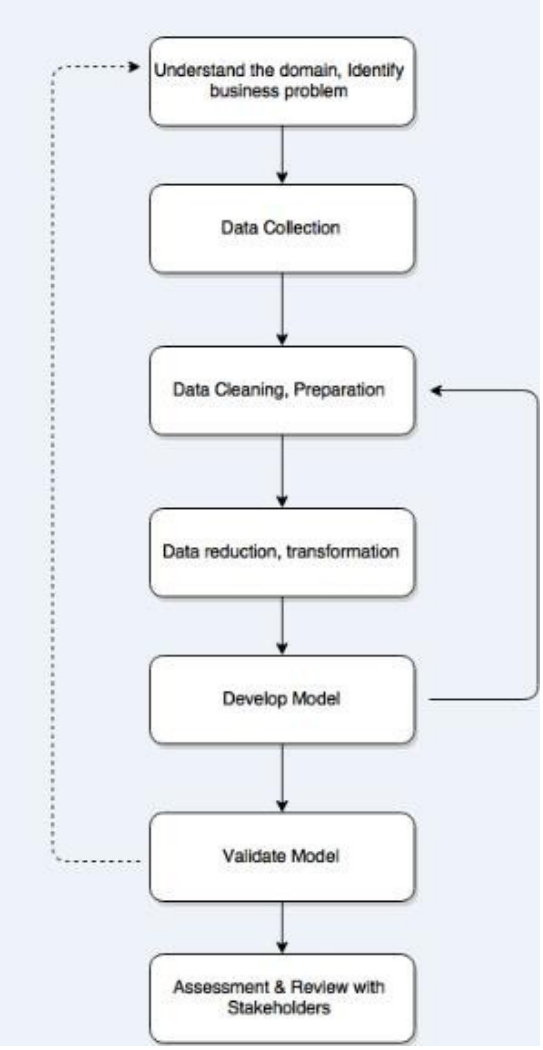
The Analytical approach will involve the following activities:

1. Data extraction from primary data source and create a robust database in postgresql
2. Data quality check

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

3. Data cleaning and data preparation
4. Study each of the variables using EDA
5. Identifying / Generating Y variable
6. Selecting the most significant variables by using combination of Correlation matrix and IV and stepwise regression.
7. Division of data into train and test
8. Model development
9. Stepwise regression and hyperparameter tuning
10. Finalizing model
11. Model validation on train and test data using Decile analysis, Gain and Lift Chart and KS Statistics
12. Verifying goodness of the model using ROC-AUC curve, confusion matrix, specificity and sensitivity checks and accuracy
13. Intervention strategies and recommendations

We plan to use the following Seven Step Analytical Approach for the Project



2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

2. Data Description and Preparation

Data Management and backend development

The data was on raiser's edge platform. The main challenge was to understand the platform, fetch the data and create a robust dataset. Created a database in PostgreSQL to replicate raiser edge's data.

As part of backend development, we created a schema and structure of whole database in PostgreSQL. Created multiple tables, views, stored procedures and functions considering machine learning model integration and PowerBI dashboard.

Data Quality

On an average 50% of the variable had missing values. Used KNN method to impute the missing values. Used US Census data to fetch income variable and integrated the same in database.

Data Preparation

Variable transformation

1. We had dataset with each donor's donation. Donor could donate more than once so primary task was to aggregate the dataset to get alumni wise records.
2. Depending on the nature of data for numeric variables, we used either logarithmic or square root transformation of such variables.
3. For Categorical data, we converted them into dummy variables based on number of unique categories.

Missing values and Outliers

1. For missing value imputation, we used KNN method.
2. For Outliers, we capped the limit as mean +/- 3 * std formula.

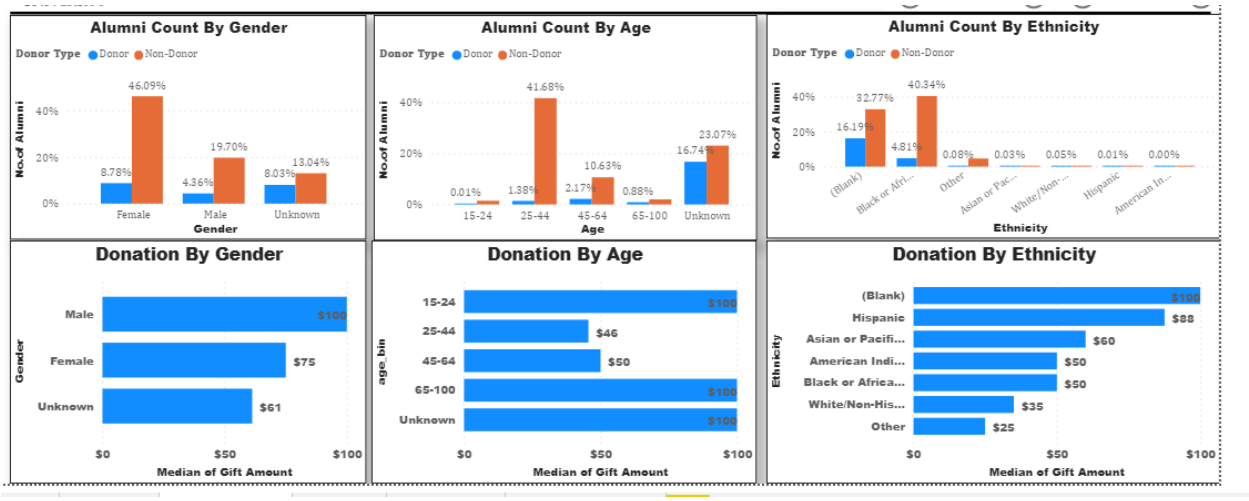
3. Exploratory Data Analysis

The exploratory data analysis is divided in major three parts. They are:

1. **Univariate analysis**: Here we use box plot / histogram / line graph etc. to check the distribution of numeric variables.

2. **Bivariate analysis**: Here we plot Bar graph to see the relationship between different continuous variables: Below graphs shows the relationship between number of visits and revenue generated:

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)



3. **Bivariate Chi-Square test**: Here we perform chi-square test to check how dependent our target variable is on various categorical variables.

Descriptive Stats:

We collect number of missing values, average, variance, standard deviation, minimum and maximum value and datapoints at different percentiles.

Data Insights and Derived variables

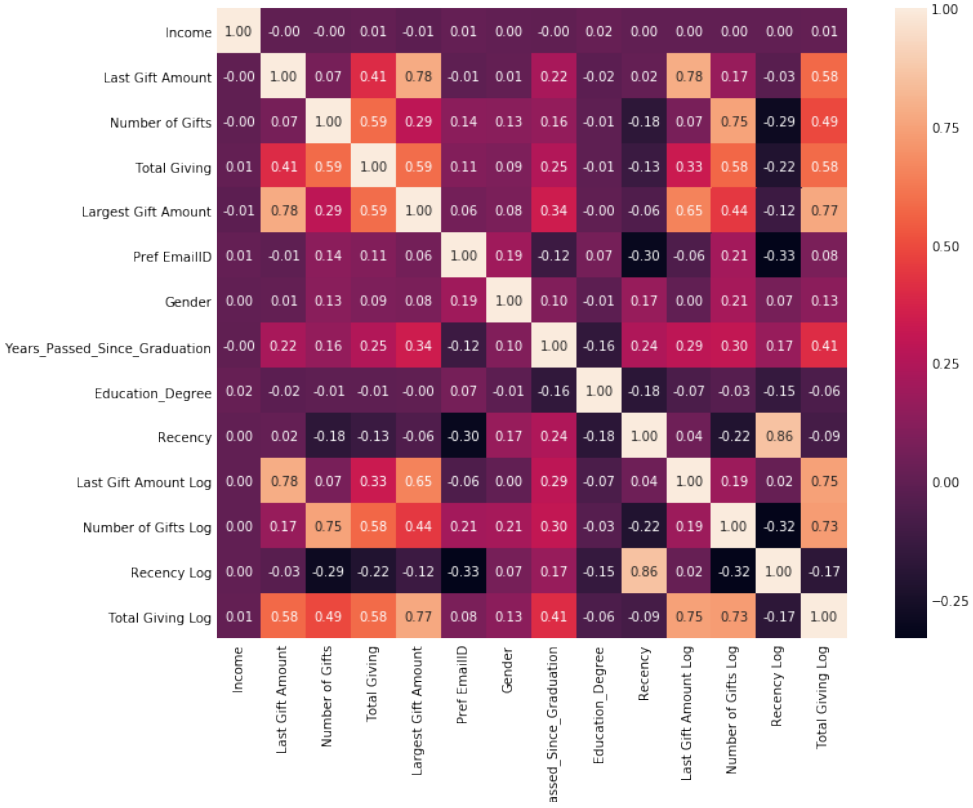
After completing EDA, we got lots of insights about the data. We could get some idea on variables that were very business specific and difficult to understand. We also felt the need to create some derived variables that might be better for final model development.

- For robust model development, it is a good practice to create derived variables and it is necessary to clean impurities in all the variables.
- We created derived variables such as age, education_degree_category, income, etc..
- There were negative age and donation values. Corrected age and donation variable by making it positive.

Correlation Matrix, WoE and IV

Correlation Matrix shows the relationship between all the continuous variable with one another. It helps to determine multicollinearity if at all exists in our dataset. Below is the correlation plot for some of the variables. Dark colour suggests negatige correlation and light colour suggests positive correlation.

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)



WOE describes the relationship between a predictive variable and a binary target variable whereas IV measures the strength of that relationship.

Using IV and correlation matrix together, we can select limited number of independent variables that are statistically significant for our target variable.

Model Development

RFM cluster analysis

Since the data was about fund raising. Recency, Frequency and monetary segmentation analysis was better to identify different segments of alumni. RFM segmentation analysis segments data considering three factors: recency(how recently alumni has donated), frequency(how many times donated) and monetary(what is the donation amount)

It helped us dividing the data into 11 different clusters(groups) each group having different characteristics than each other. Now CAU alumni relations group could target all these different groups separately.

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Logistic Regression, Decision Tree and Gradient Boosting

After data-preprocessing, we applied three different algorithms Logistic Regression, Decision tree and Gradient boosting for classification problem on overall dataset to classify alumni being donor or non-donor.

We always divide our dataset as 70% - training and 30% - validation. The model development was done at multiple levels to arrive at a most suitable model. The first one with actual variables, second one with some combination with derived variables and using different modelling techniques.

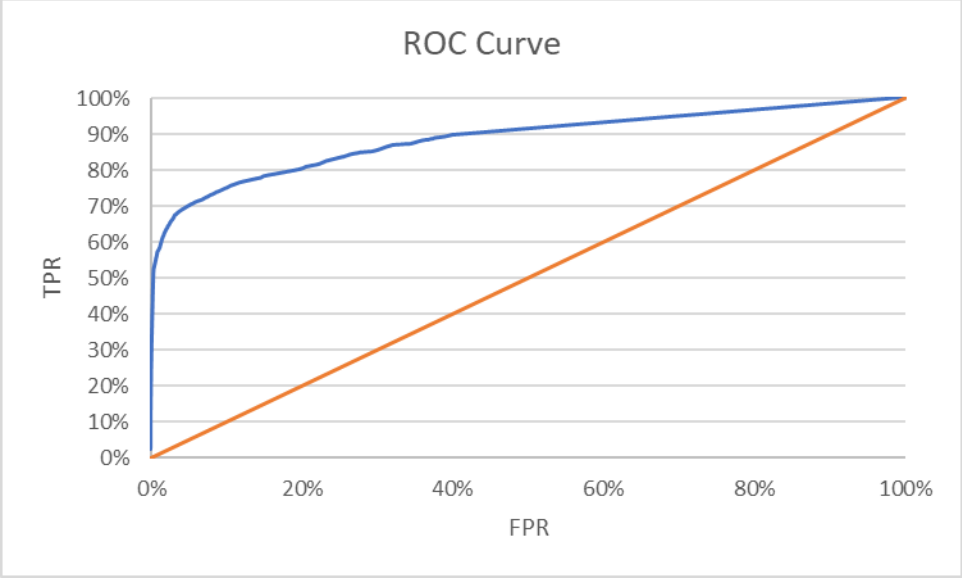
Since the objective is to predict donor or non-donor class, we used binomial logistic regression, decision tree and gradient boosting techniques.

The data for the modeling was split into two parts train & Test data. The Split of the data is as follows:

Inference:

Since we spent too much of time in variable selection, data preparation, data cleaning and initial cluster analysis exercise, we got decent result in terms of concordance / AUC despite having so many missing values.

We could achieve 0.70 AUC from Logistic regression. Below is the ROC curve for the same:



2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Validation

We used Decile analysis, Gain & Lift chart and KS statistics for initial validation of the model.

Decile Analysis:

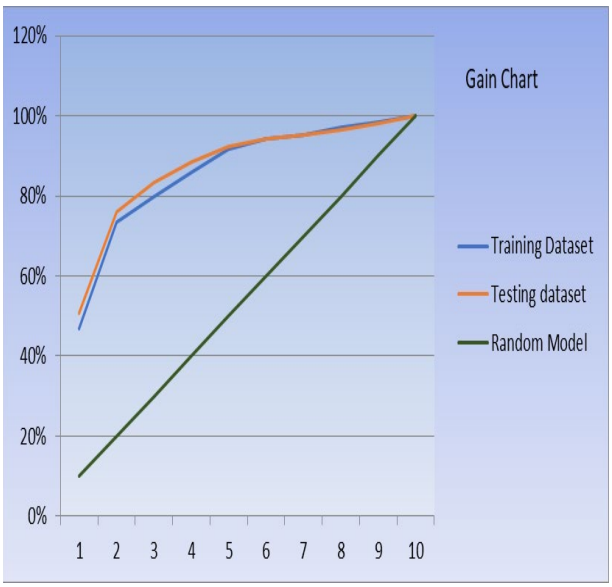
Training Dataset										
Decile	min_prob	max_prob	Non_Donor	Donor	Donor Rate	Donor %	CumDonor	Non Donor%	CumNonDonor	KS
1	72.52%	100.00%	46	489	91%	47%	47%	1%	1%	0.46085
2	20.63%	72.32%	424	277	40%	27%	74%	8%	9%	0.644992
3	11.16%	20.58%	563	65	10%	6%	80%	11%	20%	0.598619
4	7.46%	11.15%	411	63	13%	6%	86%	8%	28%	0.579702
5	5.07%	7.46%	702	62	8%	6%	92%	14%	41%	0.503582
6	4.01%	5.06%	596	26	4%	2%	94%	12%	53%	0.413366
7	3.35%	4.00%	218	10	4%	1%	95%	4%	57%	0.380839
8	3.04%	3.33%	944	20	2%	2%	97%	18%	75%	0.2176
9	2.48%	2.97%	664	13	2%	1%	98%	13%	88%	0.101754
10	0.12%	2.47%	606	16	3%	2%	100%	12%	100%	1.11E-16
			5,174	1,041						

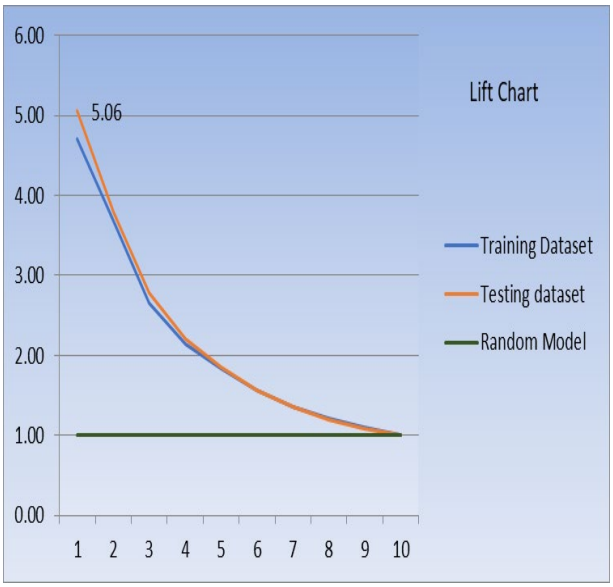
Testing dataset										
Decile	min_prob	max_prob	Non_Donor	Donor	Donor Rate	Donor %	CumDonor	Non Donor%	CumNonDonor	KS
1	73.22%	100.00%	21	222	91%	51%	51%	1%	1%	0.496257
2	19.82%	72.32%	179	111	38%	25%	76%	8%	9%	0.668655
3	11.22%	19.70%	233	33	12%	8%	83%	10%	19%	0.639106
4	7.52%	11.16%	206	22	10%	5%	88%	9%	29%	0.596636
5	5.40%	7.46%	287	18	6%	4%	92%	13%	42%	0.508649
6	4.24%	5.32%	258	8	3%	2%	94%	12%	53%	0.410918
7	3.35%	4.23%	117	4	3%	1%	95%	5%	58%	0.367445
8	3.04%	3.33%	385	5	1%	1%	96%	17%	76%	0.205801
9	2.64%	2.97%	281	7	2%	2%	98%	13%	88%	0.095454
10	0.29%	2.61%	258	9	3%	2%	100%	12%	100%	1.11E-16
			2,225	439						

RISING ANALYTICS
Data. Insight. Action.

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Gain & Lift Charts:





As we can see from Decile analysis outcome, we are getting almost similar results for our training and validation datasets. We are getting Maximum KS within first three deciles which is one of the indications of a good model. Our model is able to capture almost 76% of donors in first 3 deciles.

If we refer Gain and Lift charts, we can see that we are getting similar gain and lift for both training and testing datasets.

Confusion Matrix

After cross checking all three modeling techniques, we developed various confusion matrix based on different cut-offs. For all those cut-offs we checked various statistical parameters like precision, sensitivity, specificity and accuracy. Below is one of those examples:

TPR	56.82%
TNR	98.41%
FPR	1.59%
Precision	87.70%
Accuracy	91.47%

	Cut_Off	0.6		
		Predicted Label		
		Non-Donor	Donor	Grand Total
True Label	Non-Donor	7,281	118	7,399
	Donor	639	841	1,480
	Grand Total	7,920	959	8,879

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Model Deployment

Client wanted to calculate score of all the alumni. We implemented the whole model process in PostgreSQL to get the probability and alumni score. We implemented the whole ETL process, data cleaning, data integration from raiser's edge to PostgreSQL database.

Challenges and Recommendation

The biggest challenge was to eliminate the dependency of data access from Raiser's edge. We went through the whole backend development process from creating a database in PostgreSQL to implement machine learning model in the database itself. The second challenge was the nature of data and missing values. We had very limited set of features available so building model was very challenging.

We recommended different marketing techniques based on RFM segmentation analysis. There were many data points missing so there was also scope of data quality improvement.

Optimal Ask and Power BI Dashboard

Optimal Ask

Another request was to estimate the amount that can be asked as donation from alumni. We used two different methods to estimate optimal ask for all the alumni.

Donor Optimal Ask

We used weighted average calculation for last three occurrences of donation where more weightage was given to the latest donation amount.

Non-Donor Optimal Ask

Non-donor meant these alumni never donated to the CAU and that meant we did not have the past data to predict the optimal ask. So, we used linear regression on the first gift data for donors and derived a formula for linear regression to predict first gift amount for alumni. Implemented the same model on all the non-donors to come up with an optimal ask number. It helped CAU to get very rough estimate on what they can ask for donation from non-donors.

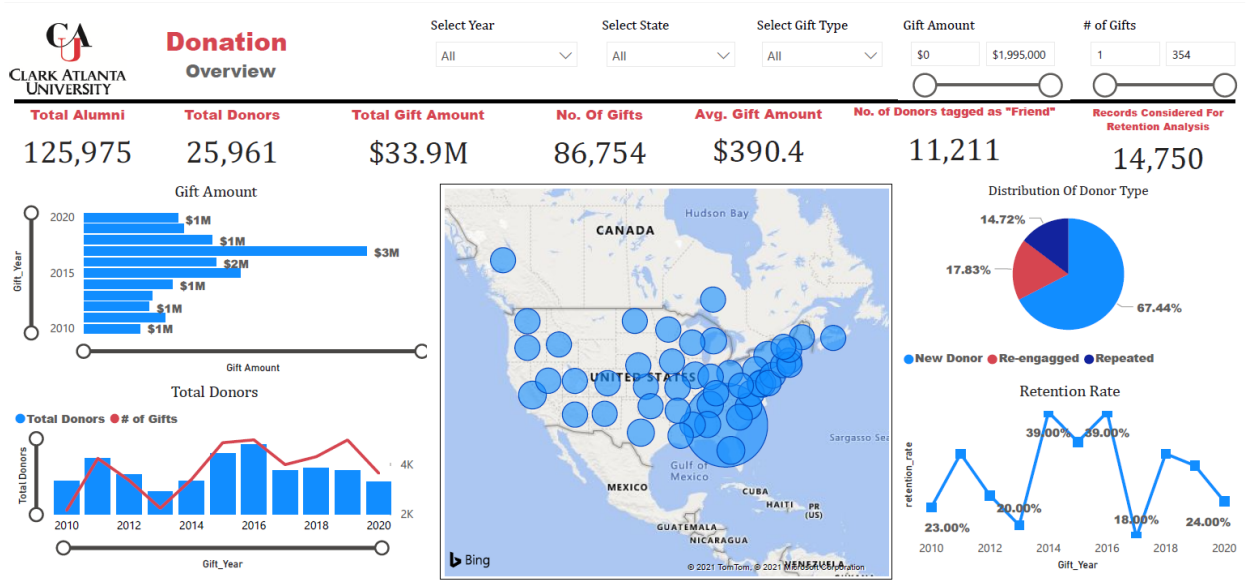
Dashboard

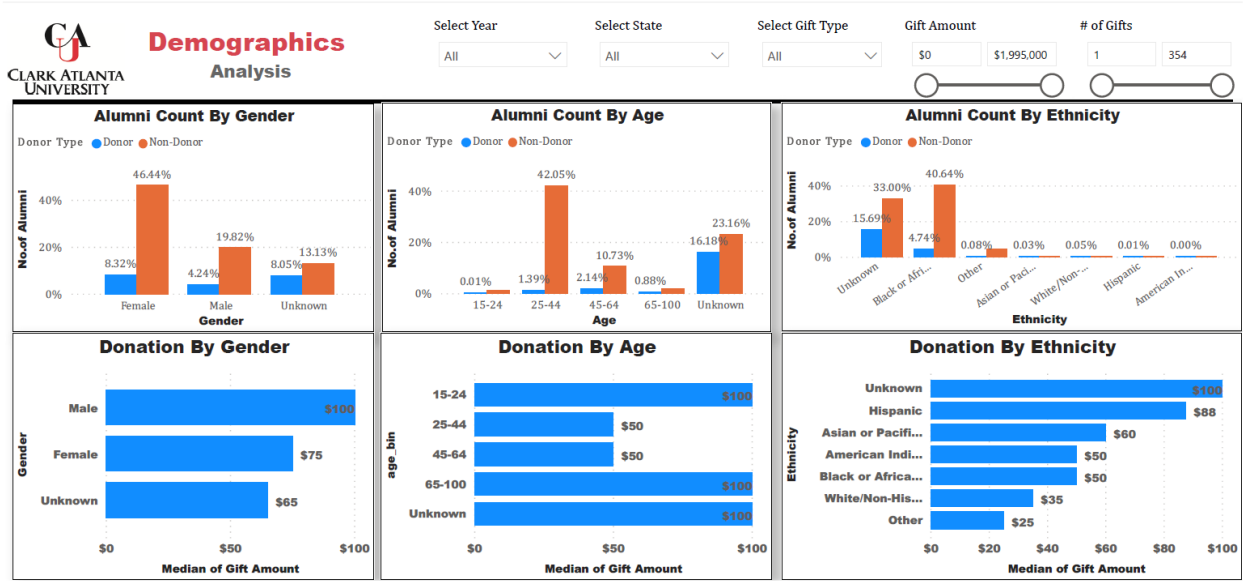
Interacting with raiser's edge and importing and exporting data was very difficult task. We came up with an idea of creating a dashboard in Power BI that can be integrated with PostgreSQL and all the basic KPIs and charts can be viewed in the dashboard.

We had to create database in such a way that the tables can be directly used to create graphs, charts and we can get KPI values very easily in Power BI. We created overview page, demographics analysis page, RFM analysis page, Search page, etc. in the dashboard itself so that the users don't have to go to raiser's edge to fetch information. Below are some snippets of the dashboard.

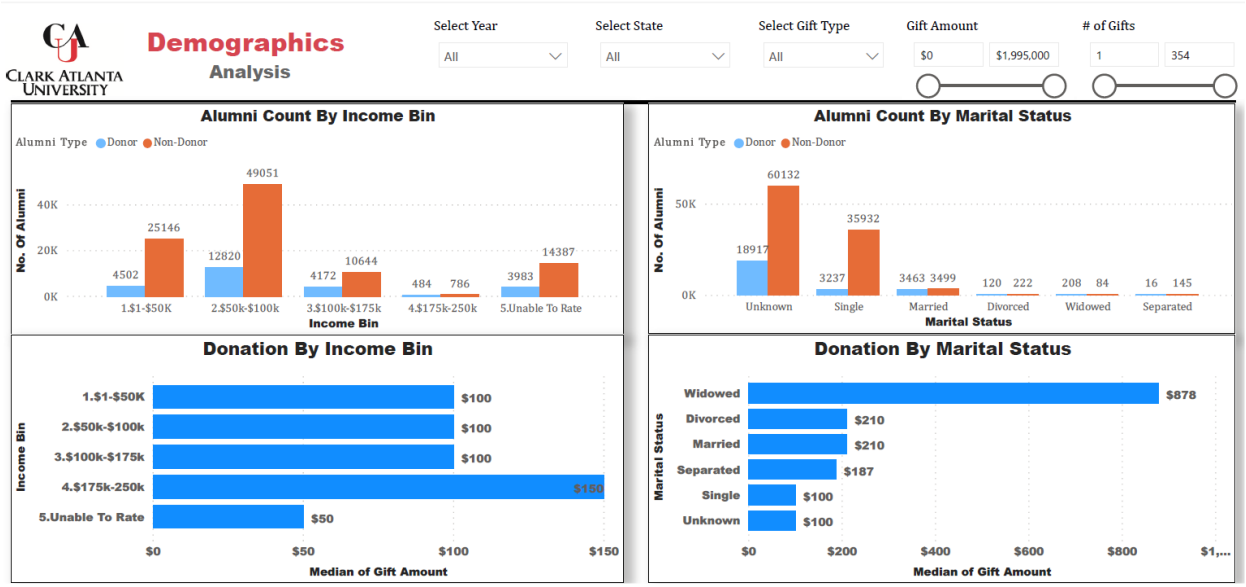
RISING ANALYTICS
Data. Insight. Action.

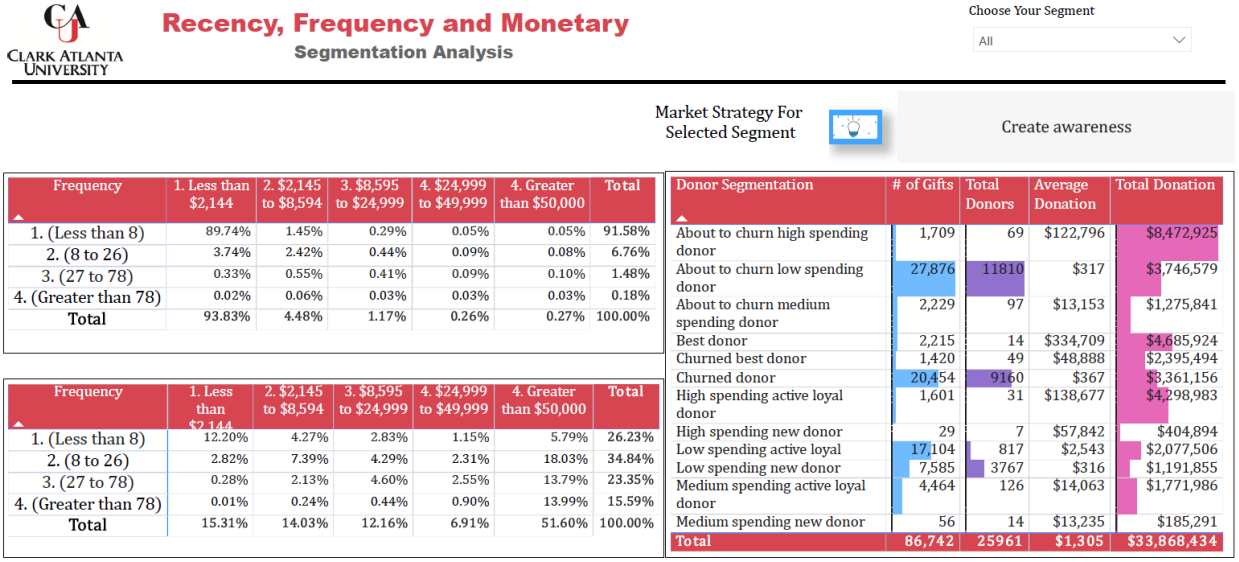
2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)





2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)





Frequency	1. Less than $2,144	2. $2,145 to $8,594	3. $8,595 to $24,999	4. $24,999 to $49,999	4. Greater than $50,000	Total
1. (Less than 8)	89.74%	1.45%	0.29%	0.05%	0.05%	91.58%
2. (8 to 26)	3.74%	2.42%	0.44%	0.09%	0.08%	6.76%
3. (27 to 78)	0.33%	0.55%	0.41%	0.09%	0.10%	1.48%
4. (Greater than 78)	0.02%	0.06%	0.03%	0.03%	0.03%	0.18%
Total	93.83%	4.48%	1.17%	0.26%	0.27%	100.00%

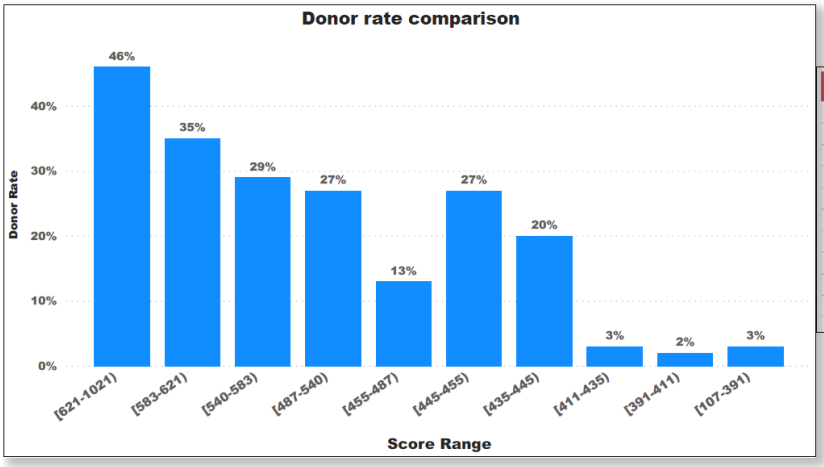
Frequency	1. Less than $2,144	2. $2,145 to $8,594	3. $8,595 to $24,999	4. $24,999 to $49,999	4. Greater than $50,000	Total
1. (Less than 8)	12.20%	4.27%	2.83%	1.15%	5.79%	26.23%
2. (8 to 26)	2.82%	7.39%	4.29%	2.31%	18.03%	34.84%
3. (27 to 78)	0.28%	2.13%	4.60%	2.55%	13.79%	23.35%
4. (Greater than 78)	0.01%	0.24%	0.44%	0.90%	13.99%	15.59%
Total	15.31%	14.03%	12.16%	6.91%	51.60%	100.00%

Donor Segmentation	# of Gifts	Total Donors	Average Donation	Total Donation
About to churn high spending donor	1,709	69	$122,796	$8,472,925
About to churn low spending donor	27,876	11810	$317	$3,746,579
About to churn medium spending donor	2,229	97	$13,153	$1,275,841
Best donor	2,215	14	$334,709	$4,685,924
Churned best donor	1,420	49	$48,888	$2,395,494
Churned donor	20,454	9160	$367	$3,361,156
High spending active loyal donor	1,601	31	$138,677	$4,298,983
High spending new donor	29	7	$57,842	$404,894
Low spending active loyal	17,104	817	$2,543	$2,077,506
Low spending new donor	7,585	3767	$316	$1,191,855
Medium spending active loyal donor	4,464	126	$14,063	$1,771,986
Medium spending new donor	56	14	$13,235	$185,291
Total	86,742	25961	$1,305	$33,868,434

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

CLARK ATLANTA UNIVERSITY

Donor Score
Analysis



Score Range	MinScore	MaxScore	Donor	NonDonor
[621-1021)	621	1,021	5,818	6,780
[583-621)	583	621	4,395	8,203
[540-583)	540	583	3,628	8,970
[487-540)	487	540	3,462	9,136
[455-487)	455	487	1,681	10,917
[445-455)	445	455	3,396	9,202
[435-445)	435	445	2,483	10,114
[411-435)	411	435	420	12,177
[391-411)	391	411	260	12,337
[107-391)	107	391	418	12,179

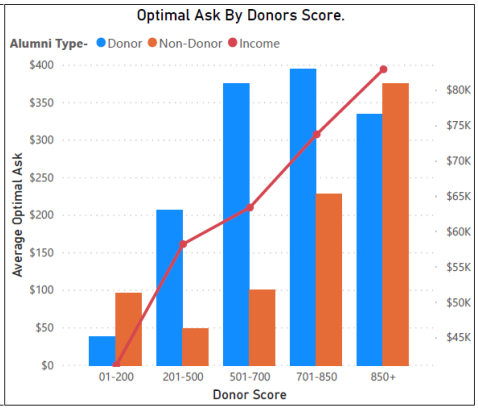
CLARK ATLANTA UNIVERSITY

Optimal Ask
View

Select Degree: All
Select State: All
Age Bin: All
Donor Score Bin: All
Optimal Ask: $1 – $999,030

Search by Name:- All

Name	Age	Optimal Ask	Income	Donor_Score	Primary Education Degree	State
Zytia Simmons	31	$37	$41,392	411	Unknown	Florida
Zykilya Sizemore	29	$17	$48,382	390	Unknown	Georgia
Zykia Stewart	27	$24	$35,065	380	Unknown	Louisiana
Zykia Spencer	29	$18	$30,392	453	Unknown	Georgia
Zyed Kane	28	$17	$51,603	419	Unknown	Illinois
Zyairra Alexander	26	$22	$0	372	Unknown	
Zwella Harris	46	$99	$51,367	529	Bachelor of Arts	Georgia
Zuri-Starr Turner	37	$33	$58,735	277	Master of Social Work	California
Zurika Knox	29	$17	$32,848	398	Unknown	Michigan
Zuri Sullivan	34	$30	$63,580	477	Unknown	Florida
Zuri Kennedy	28	$15	$80,325	408	Unknown	Georgia
Zuri Harris	32	$82	$104,048	443	Bachelor of Arts	Tennessee
Zuri Brooks	42	$109	$40,192	535	Bachelor of Arts	Louisiana
Zuri Ali	27	$17	$79,522	454	Unknown	Illinois
Zuri Ali	28	$19	$38,592	461	Unknown	Ohio
Zuri Agbaje	27	$28	$54,225	435	Unknown	New York
Zuri Adili	46	$61	$114,692	470	Master of Arts	Georgia
Zuleyma Sois	25	$43	$53,103	428	Unknown	California



2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)