Wells Fargo Data Analytics: http://projects.miscthings.xyz/wellsfargo

Deliverable 1:

Analytical Process: Figure 1

The most important stepping stone to consider when starting the analytical process is the business objective. The objective stated in the challenge was to determine which outreach methods yield the best product portfolios. The superlative best in this regard was defined to be the size and the growth of the portfolio during the 12 month period. The specific insights to achieve this objective that were given are to determine factors of growth, determine if demographics are correlated, and determine how customer interaction through outreach types and channels correlate with the customer's portfolio. Thus, the business objective inherited for this analytical process is to determine which variables correlate, then to determine a model to map a growth model either monthly or yearly that holds covariance or assumes independence. This end goal in the analytical process matches the original business objective because it takes into account all variable correlations to piece together healthy customer interactions which could be interrelated or purely independent.

The first step of the analytical process flow is to run simple single variable statistics on the columns themselves. In this step, the goal is to determine the total sum, mean, variance, and standard deviation of each variable besides customer number and month. The total sum will determine if any method was used significantly more than another during the year for each customer, which in turn is also assisted by the mean which shows the average per customer in the dataset. An important distinction to make is that the mean is measured monthly, and thus to determine average interactions in the entire period it must be multiplied by 12. The variance and standard deviation show whether a similar strategy was employed for each customer or whether it greatly varied depending on each customer. Understanding the total times a method was implemented, and whether it was implemented evenly across the dataset or not will be able to shed a light on the results we find later on and make sure to dismiss any biases in the information this analytical process finds.

The next step is to run covariance and correlation tests between each variable. The most important metric in this step is the correlation which is derived using the correlation coefficient which takes their covariance and scales it off each variables standard deviation. The correlation will allow us to see if more customer outreach of a certain type is either positively or negatively correlated with account balance, account numbers in category A or B, or product types. Besides being useful for checking metrics of the portfolio, the correlation will also be able to show how overall strategies played out in the dataset. For example, if a certain channel was use more often with an outreach type, then it will have a positive correlation.

Therefore, with single variable statistics and correlation complete, it is then time to process the information into tables to determine which methodologies should be employed.

Results from Exploration of the Data: Figure 2

After the preliminary analysis, it is found that there are outreach types that have much larger variance than others, for example cust_outreach_avi has the largest sum with the largest variance in the dataset. Furthermore, there are less and lower in magnitude negative correlations in the dataset that only reach -0.00937359. While there are strong positive correlations in the dataset

for both metrics and methods of outreach such as wf_outreach_flag_chan_i and wf_outreach_flag_chan_ii have a correlation of 0.520937 while wf_outreach_flag_chan_ii and typeB_ct have a correlation of 0.241825. Thus in order to achieve the business objective it is imperative to determine what positively correlates with growth on different time scales depending on methods used.

Methodologies: Figure 3 & 4
The first method used to further the analytical process was to create a linear regression model based on monthly growth percentage in the data based on different categories versus the important metrics of balance, account numbers, and product flags. These metrics determine the best portfolio growth as it portrays the size of the balance, the number of accounts, and the diversity of the products in the portfolio. The linear regression assumes an independence between each month in order to ensure that one month's actions do not affect the other months' actions. In addition, the formula uses percent growth to scale the growth based on the account rather than simply the magnitude which could skew the data. Therefore, this method creates a monthly model to suggest the best possible method to increase growth for a single month. The data for this methodology will be portrayed as linear graphs that have assumed independence.

The next logical step from the independent monthly model is to extrapolate the data to hold a persistent yearly model. This model will take into account the correlation between the metrics and the categories while layering it on top of the entire year's performance. This will take into account the customers who had persistent events occurring versus those who had one time events, and thus it shows how to better handle customers, either with constant customer interaction or intermittent interaction. The data for this methodology will be portrayed in bubble graphs which accurately scale with growth levels by percentage in a 3 dimensional format that compares correlations between demographics, outreach types, channels, and the metrics.

Excluded Methods:
Methods that did not make the cut are some machine learning algorithms that are used in classification cases. Since the business objective was to derive what drives growth, it was more imperative that the methods employed discovered what conditions growth occurred in before starting to classify growth groups. Such algorithms are SVM, KNN, Naïve Bayes, and K-Means. SVM, KNN, and K-Means, are effective in determining classifications, but would also suffer under the n-curse dimensionality because there were a myriad of categories and variables, and thus the space used to classify in these algorithms would be too large. Bayes is tempting to employ a probabilistic model to determine growth; however, the method uses strong assumptions of independence which were deemed to be inappropriate in this setting.

Deliverable 2: Figure 1,2,3,&4
Code Composition:
The code used for this challenege is mainly in PHP for the analyzing portion. The data is loaded into a MySQL database using imports from the excel file that was converted into a csv format, then formatted into sql import using regex statements. The dynamic graphing is created by Chart JS library which can be found here: http://www.chartjs.org/ . The fetching of the data form the server is handled by JQuery and JavaScript. The main HTML of the website is formatted for

BootStrap v3. This concludes the langauge composition of the code which can also be viewed in the GitHub repository for the code. The GitHub repository is hosted here: https://github.com/ZiBuDo/WellsFargoAnalytics .

The architecture of the code resembles most websites with a css/ directory fo css style sheets, js/ directory for JavaScript files, assets/ directory which contains php/ directory for PHP scripts and sql/ for the sql imports that are derived from the data analysis, and img/ for the figures to be loaded onto the web page.

Analyze.php:
The main PHP scripts used for the data analysis is analyze.php and fetch.php. Analyze.php is used for the majority of the analytical flow. Comments in the code which tell you what the next loops or functions purpose and action are can be found through out the code such as : // Calculate basic stats for all areas such as mean, std. deviation, variance, sum. This program loads the dataset into memory on a server to make the process extremely quick.

The first step, single variable statistics is handle in the first loop of the code that loops through all 120,000 records twice to calculate sum, mean, variance, and standard deviation. These are then entered into a SQL table called WF1VARSTATS on the server for the fetch.php to show to the web page.

The second step, two variable statistics uses the next double foreach loop. This loop goes through each variable and compares it to each column to find their covariance. Once the covariance is found it is entered into the SQL table WF2VARSTATS and used to find the correlation coefficient. The correlation coefficient is found by dividing the covariance and dividing it by the standard deviation by each variable. This yields a proper measurement of correlation which is entered into the same table to be fetched by fetch.php.

Fetch.php:
The exploration step was used by implementing fetch.php. Fetch.php interacts with index.js. Index JavaScript sends AJAX GET requests to the server which fetch.php handles. Fetch.php handles the request by checking what the request is for, then delivering html content or JavaScript objects to be show as graphs or tables. In order to explore the first two steps to decide on the models to be used, the data was placed into two graphs with red and green highlighting for forrelation, see Figure 2.

Models:
To build the linear regression independent monthly gross model, there was a linear regression function created in analyze.php which could take x and y arrays, or two variables and determine the slope and intercept of the line. The two variable groups are the categories of the variables and the measurement variables named deltas. The categories were mainly demographics, balance categories, customer outreach attempts, and outreach channels. The metrics that were tracked were the balance, account count for A and B, and product flags. The linear regression is calculated by finding the total percent growth each month for each metric as the y variable while the x variable is category. These are then graphed as show in figure 3 which shows % growth in a month based on the category's value.

Finally, the persistent growth model is created in the rest of the program. The next 3 loops determine demographic bubble chart data by creating JavaScript Data Objects to pass to the front-end to graph in Chart JS. Bubble graphs are chosen for this model because it effectively portrays correlation between two variables by graphing them on the x and y plane, but then scales the positive or negative growth through the radius of the data point. For the demographics, these graphs are shown with demographic B on the y-axis and demographic A on the x-axis with a different metric tracked depending on the selection in the tool. The radius of the circles are scaled by the raw pixel size from 5 to 25 which is scaled by a % growth based on the minimum and maximum values of the growth in the year for that metric with regard to 0.

The final charts that are created under this model are those that map customer outreach types versus outreach channels in figure 4. These also vary by the metric; however, the user is able to use the tool to map each outreach type with each out reach channel for a total of 32 possible x and y combinations and then try each mesurement of balance, account count A, and account count B.

The PHP scripts are in assets/php/, these scripts managed the sql import and the creation of the data for these tables and graphs to create over 250 graphs and tables to analyze the data at hand.

Deliverable 3:

Growth: Figure 3,5,6,7 & 8
Question: What drives growth in accounts and/or balance between month 0 and month 12?

This question is best answered through the single and double variable statistics because it shows overall large trends that drive growth in accounts based on account balance, account diversity with regard to products, and account growth.

Account growth in terms of balance did not correlate highly positive or highly negative with any other column to be noticeably different as seen in figure 5. However, the number of accounts A and B both highly correlate positively with outreach channel i, ii, and iv; they also correlated positively with outreach type ai, aii, aiii, and avi. See figures 6 and 7.

Therefore, growth by number of accounts is highly correlated to certain outreach channels and outreach types. On the other hand correlations found for product diversity based on product flags A-E correlated with outreach types, but with a smaller magnitude, and thus it is less likely to occur within the dataset. See figure 8.

These findings can also be corroborated with the independent growth model by monthly growth percentage. For example figure 3 shows cust_outreach_aii is highly correlated with balance. To figure out how much growth a column can yield, check it out on the independent growth tool section of the web page.

Demographics: Figure 9

Question: What demographic types, if any, are more likely to increase (or reduce) their number of accounts and/or balance between month 0 and month 12?

The demographics that are more likely to increase their balance or accounts are those with larger green circles in figures 9. The demographics that are more likely to decrease their balance of the number of accounts are those with larger red circles in figure 9. Small green or small red circles means those are negligible differences; however, those demographics still are more likely to decrease than to increase or vice versa depending on the color. Demographic A is cust_demographics_ai, while Demographic B is cust_demographics_aii.

Notable demographics that would increase account balance or accounts are, given as ordered pair with (A, B) demographics: (0,2), (1,3), (2,1), (2,5), and (3,4).
Notable demographics that would decrease account balance or accounts are, given as ordered pair with (A, B) demographics: (0,3), (1,5), and (4,4).

Customer Interaction: Figure 4
Question: What types of accounts, customer interactions, customer events, or Wells Fargo outreach, are more correlated with account and/or balance change.

The customer interactions, customer events, and outreach methods that was most correlated with account and balance growth can be derived from the persistent growth model which tracks the outreach type versus the outreach channel based on account numbers or balance by % growth per year. This model takes into account all interactions the customer had throughout the year based on different channels and portrays how it affected growth by each metric. See figure 4.

These graphs have a pattern where the outreach channel has most growth either close to 1 or closer to 0 which means that no matter which outreach type the customer is dealing with, with regard to the outreach channel then it matters a lot that the customer has a consistent experience where the channel is either consistently used or not used. Most large losses appear in between .4 and .7 in the outreach channel which means the outreach is occurring between 40 to 70 percent of the time in that channel rather than closer to 80% of the time. With regard to outreach type, the lower the number the better which means that the lower number of outreach of each type is more effective than a large amount of outreach types. Most large losses are seen when outreach type exceeds 2 standard deviations.

Conclusion:
The analysis found within the tool and the graphs shows strong trends in correlations between demographics and customer interactions with regard to increase in growth in the portfolio. Each model was created with the each of these questions in this deliverable in mind when exploring the data, and it appears that the models have shown where significant growth or losses appear in all three questions. Business wise, this tool is quite important because it shows how to carry out customer interaction in the future, and which customers to focus with different tactics to maximize growth, while avoiding reducing balance sizes and account numbers. The models available take into account both independence at a time interval, but keep the correlation between columns, and thus creates the most accurate picture for the dataset. For further analysis, trends can be classified into high growth, medium growth, and low growth, and the same with

probability of growth utilizing SVM and KNN machine learning algorithms which should flourish now that the significant correlated columns have been identified, and thus have removed n-curse dimensionality.

Figures:

Figure 1:

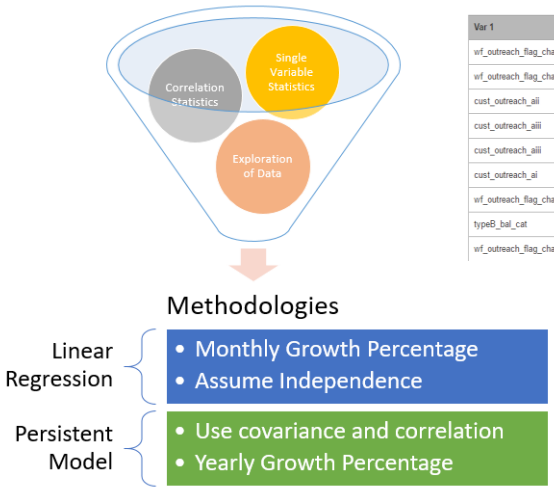

Figure 2:

Var 1	Var 2	Correlation
wf_outreach_flag_chan_i	wf_outreach_flag_chan_ii	0.520937
wf_outreach_flag_chan_ii	wf_outreach_flag_chan_i	0.520937
cust_outreach_aii	cust_outreach_aiii	0.518843
cust_outreach_aiii	cust_outreach_aii	0.518843
cust_outreach_aiii	cust_outreach_ai	0.42179
cust_outreach_ai	cust_outreach_aiii	0.42179
wf_outreach_flag_chan_ii	typeB_bal_cat	0.414531
typeB_bal_cat	wf_outreach_flag_chan_ii	0.414531
wf_outreach_flag_chan_ii	typeA_bal_cat	0.392923

Figure 3:

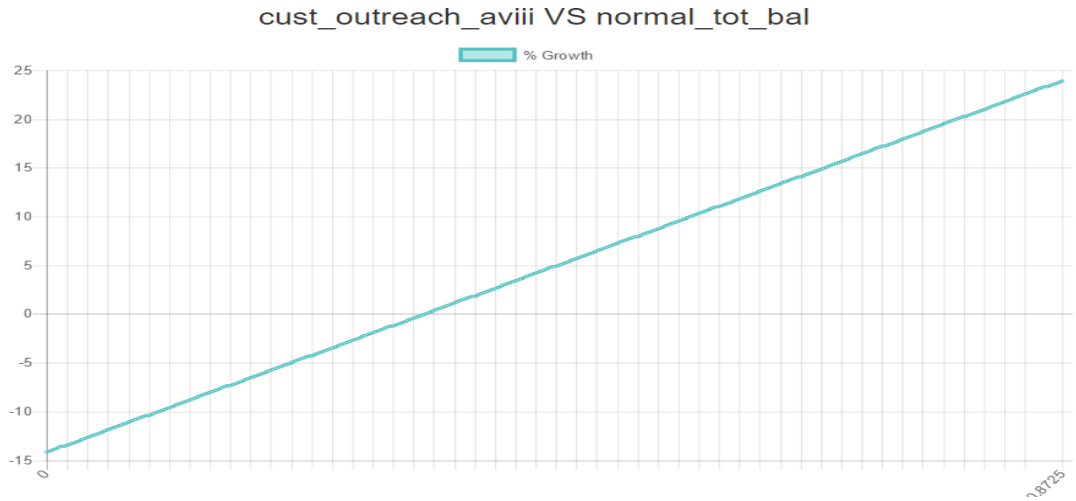

Figure 4:

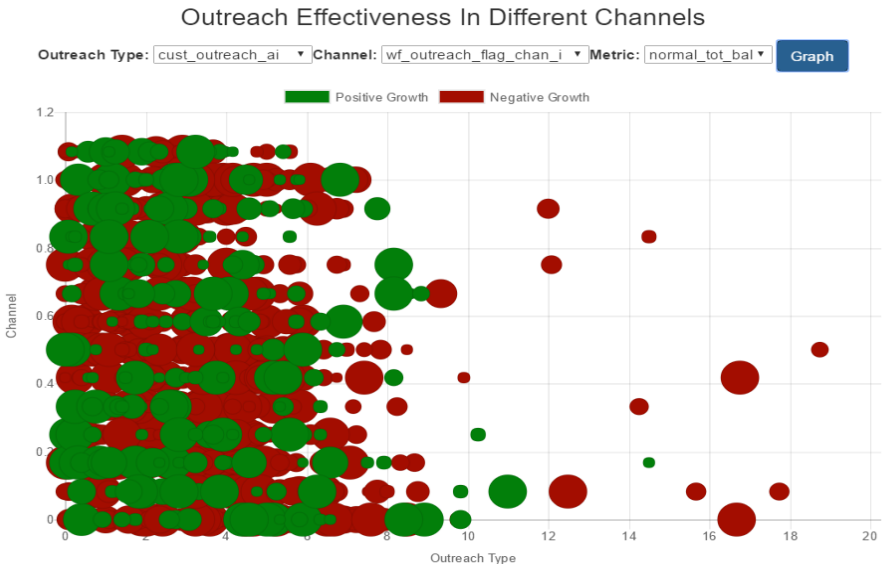

Figure 5:

VAR1	VAR2	STAT	VALUE
normal_tot_bal	typeB_bal_cat	CORRELATION	0.0240612
normal_tot_bal	typeA_bal_cat	CORRELATION	0.0194792
normal_tot_bal	cust_outreach_aiii	CORRELATION	0.0110353
normal_tot_bal	cust_outreach_ai	CORRELATION	0.0107365
normal_tot_bal	typeA_ct	CORRELATION	0.00859734
normal_tot_bal	typeB_ct	CORRELATION	0.00606398
normal_tot_bal	typeC_bal_cat	CORRELATION	0.00514424
normal_tot_bal	cust_outreach_aii	CORRELATION	0.00463888
normal_tot_bal	cust_demographics_aii	CORRELATION	0.00381831
normal_tot_bal	cust_demographics_ai	CORRELATION	0.00364402
normal_tot_bal	typeD_flag	CORRELATION	0.00332632
normal_tot_bal	typeD_bal_cat	CORRELATION	0.00251214
normal_tot_bal	cust_outreach_aiv	CORRELATION	0.0024686
normal_tot_bal	typeC_flag	CORRELATION	0.000920224
normal_tot_bal	typeG_flag	CORRELATION	0.000246479
normal_tot_bal	wf_outreach_flag_chan_iii	CORRELATION	0

Figure 6:

VAR1	VAR2	STAT	VALUE
typeA_ct	wf_outreach_flag_chan_ii	CORRELATION	0.216317
typeA_ct	typeA_bal_cat	CORRELATION	0.197097
typeA_ct	typeB_bal_cat	CORRELATION	0.186023
typeA_ct	wf_outreach_flag_chan_i	CORRELATION	0.151458
typeA_ct	cust_outreach_avi	CORRELATION	0.147112
typeA_ct	cust_outreach_aiii	CORRELATION	0.144278
typeA_ct	cust_outreach_ai	CORRELATION	0.12781
typeA_ct	typeB_ct	CORRELATION	0.12184
typeA_ct	typeF_flag	CORRELATION	0.121733
typeA_ct	wf_outreach_flag_chan_iv	CORRELATION	0.114029
typeA_ct	cust_outreach_aii	CORRELATION	0.112896

Figure 7:

VAR1	VAR2	STAT	VALUE
typeB_ct	wf_outreach_flag_chan_ii	CORRELATION	0.241825
typeB_ct	typeB_bal_cat	CORRELATION	0.212212
typeB_ct	typeA_bal_cat	CORRELATION	0.191125
typeB_ct	wf_outreach_flag_chan_i	CORRELATION	0.171338
typeB_ct	cust_outreach_avi	CORRELATION	0.153282
typeB_ct	cust_outreach_aiii	CORRELATION	0.151237
typeB_ct	cust_outreach_ai	CORRELATION	0.133233
typeB_ct	typeF_flag	CORRELATION	0.128269
typeB_ct	wf_outreach_flag_chan_iv	CORRELATION	0.126634
typeB_ct	typeA_ct	CORRELATION	0.12184
typeB_ct	cust_outreach_aii	CORRELATION	0.121407

Figure 8:

VAR1	VAR2	STAT	VALUE
typeD_flag	typeD_bal_cat	CORRELATION	0.313627
typeD_flag	wf_outreach_flag_chan_i	CORRELATION	0.0887217
typeD_flag	typeC_bal_cat	CORRELATION	0.0851092
typeD_flag	wf_outreach_flag_chan_ii	CORRELATION	0.0764003
typeD_flag	typeA_bal_cat	CORRELATION	0.0762097
typeD_flag	typeB_bal_cat	CORRELATION	0.0694853
typeD_flag	cust_outreach_aiii	CORRELATION	0.0624625
typeD_flag	cust_outreach_avi	CORRELATION	0.0609175
typeD_flag	cust_outreach_ai	CORRELATION	0.058816
typeD_flag	typeC_flag	CORRELATION	0.0584294
typeD_flag	typeA_ct	CORRELATION	0.0524928
typeD_flag	typeF_flag	CORRELATION	0.0509023
typeD_flag	typeB_ct	CORRELATION	0.0480171

Figure 9: