EASTERN UNIVERSITY

Project Overview

DTSC691: Applied Data Science

Sales Standards of Excellence (SSE)

Diana M. Bowden
Fall Term 2023

Contents

Project Goal

The purpose of this project is to provide information to sales leadership to model the most successful sales representatives, which serves as a foundation for future metric activity quotas and overall best-selling practices. In other words, if a sales leader knew the predictors for sales representative success (100% or more quota attainment), they could modify weekly minimum metric requirements based on the predictors and their corresponding metrics.

The project structure is based on the organization's Enterprise level, cloud-based computing platform Microsoft Azure. On Azure, the Company utilizes Apache Spark, which is a system for executing engineering, data science, and machine learning all on one platform, using single-node machines or clusters. Apache Spark is a framework for executing code in parallel across many different machines. Databricks is an API that runs on Apache Spark which offers several languages including SQL, Python, R, an integrated visualization tool and Machine Learning which is all housed on the same platform.(databricks.com)

The Relias CRM (Customer Relationship Management) is Salesforce. The Salesforce (SFDC) tables and data are available within the Databricks platform. The scope of the SFDC data spans 10 years across our various sales divisions (Account Management, Small Business, Mid-Market and Enterprise) as well as verticals served (post-acute, acute, and health and human services). Thresholds for each sales division as follows:

- Small Business (SMB): This sales team focuses on organizations with 10-199 employees.
- Mid-Market (MM): This sales team focuses on organizations with 200-1999 employees.
- Enterprise (ENT): This sales team focuses on organizations with 2000+ employees.
- Account Management (AM): This sales team focuses on current customers, in the SMB, MM and ENT bands.

Data

SFDC objects were used as the data sources. Relias integrates many operational, product, vertical, legal, implementation, financial and activity metrics all within Salesforce, therefore all data relevant to this project has been available through Salesforce. Forty-three SFDC tables and their fields were analyzed to determine where the target data resides.

Confidentiality Disclosure

Since this project contains real data from my employer, PII (Personal Identifiable Information) is present throughout the source code. I have therefore minimized certain cells containing PII within the notebooks and final dataframes, to prevent PII disclosure. In the video presentation as well as the source code documentation, cells were also minimized to prevent disclosure of PII.

Data Exploration and Software Selection Changes:

The following ten out of the 43 SFDC tables were identified and explored in greater detail during the data evaluation process. After selecting the top nine tables that contain the desired data, exploration of the data in SQL was conducted.

1. Opportunity
2. Account
3. Task
4. Rubybenchmarking
5. Servicecontract
6. Opportunitylineitem
7. Sbqq__quote__c
8. Forecastingquota
9. Bookingsbudget
10. Opportunityhistory

Software Exception - While Data Bricks provides an API that integrates SQL, Python, R and Machine Learning, I unexpectedly found that during the data exploration process, my traditional Python coding used in NumPy and Pandas did not work. While the concepts were the same, the actual coding was often different. I used the pandas API within Databricks initially, however, Spark was needed for computing power. Also, I initially used the wildcard conversion tool from Spark to SQL, however, I learned to use Spark SQL for data exploration.

During the initial data exploration, I ran queries on data formats to determine changes required. Descriptive statistics as well as visualizations, such as histograms, scatterplots, and boxplots, were created to aid in data comprehensive, distribution, identification of outliers, Naan values and potential correlations. The integrated visualization tool within Databricks was leveraged to conduct these visualizations. A few of these visualizations are captured in the following figures.

Figure 1. Histogram of Average Number of Products Sold per Opportunity – All sales divisions.

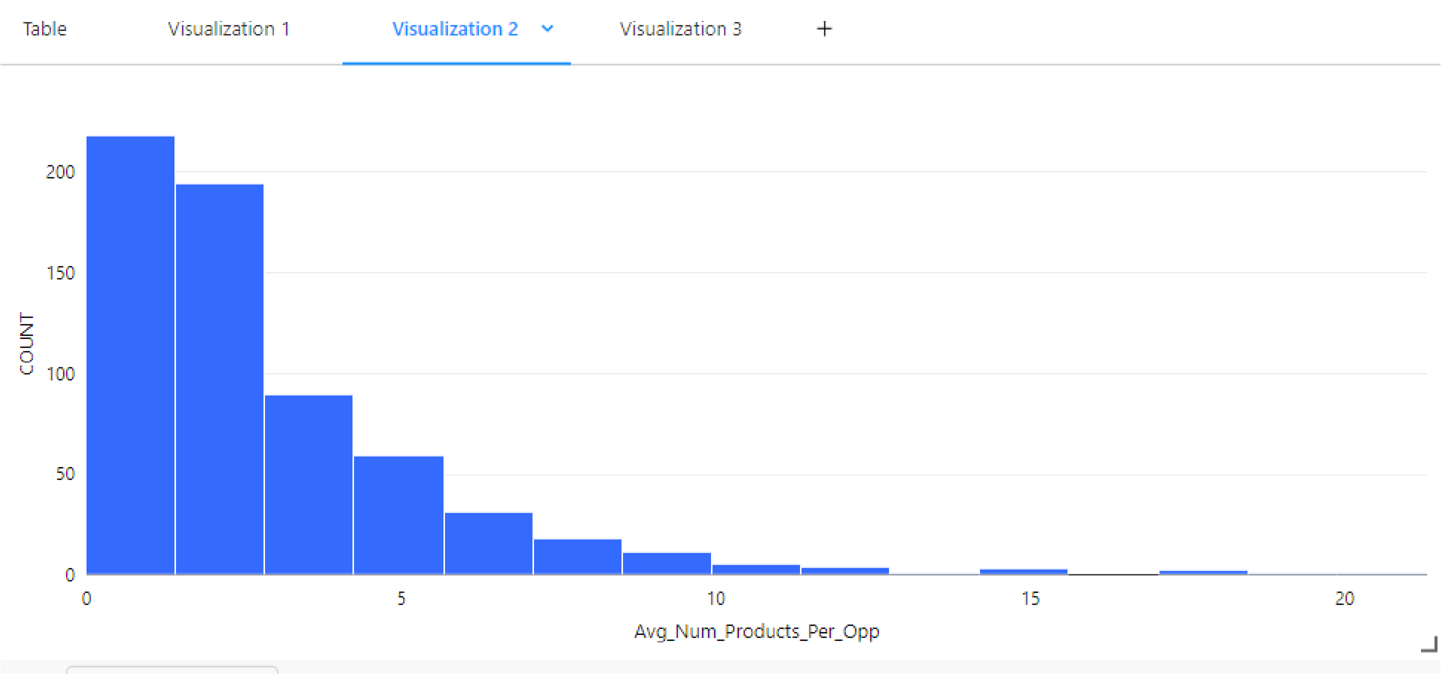


Figure 2. Scatterplot: Relationship between Average Opportunity Amount and Percent To Quota -All sales divisions

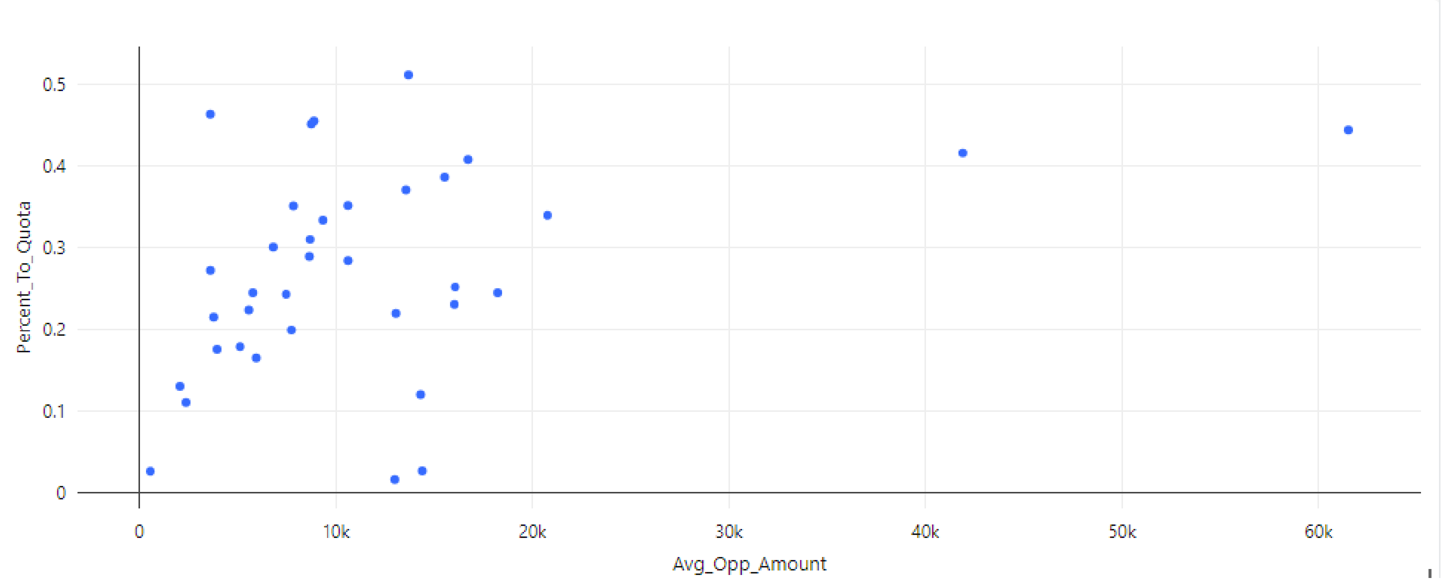
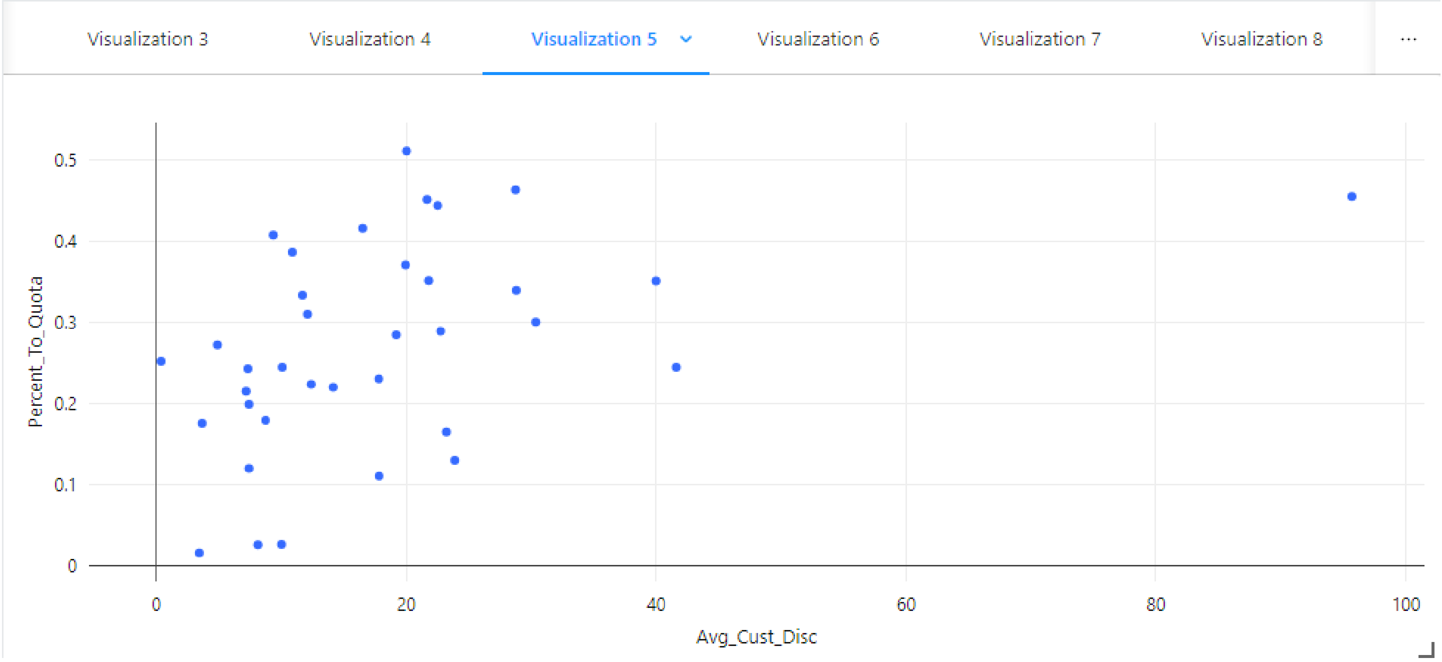


Figure 3. Scatterplot of relationship between Average Customer Discount and Percent To Quota-SMB



Exploration Findings

Notes were taken on data wrangling requirements including, but not limited to, the following checklist:

- Timestamp date – this needed to be converted to a regular Date Field
- Removal of Account Manager contract renewals, which skewed the data.
- Naan values were notated in various fields, including the dependent variable.
- A list of calculated fields required for the regression models were identified.
    - Rep Start Date, Rep term date.
    - Min and max activity dates to determine days worked for each rep.
    - Days Worked
    - Weeks worked –This data was important in normalizing sales data over their career history so that averages were calculated by week, regardless of rep tenure.(it is important to note that I did look at rep tenure as potential criteria / feature in the model)
    - Tasks, broken up by each type including email, demos set, demos performed, demo requests, calls, conversations, online meetings.
        - Total online meetings (total and per rep per week)
        - Total emails (total and per rep per week)
        - Total customer calls (total and per rep per week)
        - Total prospect calls (total and per rep per week)
        - Total conversations (total and per rep per week)
        - Total demo request (total and per rep per week)
        - Total demos set (total and per rep per week)
        - Total demos performed (total and per rep per week)
        - Total activities (total and per rep per week)
        - Total opportunity amount per rep
        - Total accounts

- The “IsWon” field was Boolean therefore this needed to be broken down and aggregated to understand the number of closed/won vs. closed/lost
    - Form this data, I was able to create a calculated field of close rate (% of deals that a sales rep closed)
- Average discount
- Average Opportunity amount by sales division
- Total opportunity amount by rep
- Average term length (contract term)
- Average Sell Cycle
- Average number of products sold per opportunity
- Sum of quota
- Percent to quota – Dependent Variable

Data Preparation and Cleaning

Originally, Python was supposed to perform any required data cleaning and wrangling.

Exceptions – After working on the Databricks platform, I found that Python Pandas and NumPy did not work. I pivoted and learned to use Spark (PySpark) which is complimentary to Python, though different. Spark has a Python API for Apache Spark which helps interface with Resilient Distributed Datasets (RDDs). (www.databricks.com)

Strategy Change

Joins

In addition to some Spark coding challenges, another significant issue I encountered during the data cleaning was my strategy. For example, in my first pass of joining nine selected tables, I successfully created a data frame to clean and prepare. However, when I started creating calculated fields, the strategy was unsuccessful. This is because the aggregations needed to be performed in separate data frames before joining them. One of the issues was the millions of rows of activity data that I had not anticipated which needed to classify, cleaned, and aggregated before the joins.

To better understand the data and relationships to one another, as well as the order of the joins, I created a schematic of the desired data frame that corresponded to each of the tables I was to join. From this schematic, I was able to clean the data and add the calculated fields to each segment of the data frame before joining them together (please see diagram one in the appendix for the mockup diagram).

Aggregation of Sales Divisions

To increase the sample size, I opted to initially pool the sales division data. To account for the differences among divisions, I normalized the quotas, using a “Percent to Quota” metric as well as their tenure (weeks worked).

Correlations

The most promising data correlations were first determined through visualizations (histograms and scatterplots) before running Pearson’s correlations as follows:

***Table 1. Experiment A: Pearson’s Correlations on the first set of features*.**

Feature	**Dependent Variable: Percent To Quota (0 to 1)**
Avg_Opp_Amount	0.8070
Avg_Sell_Cycle	0.2990
Total_Demo_Request_Per_Week	-0.0300
Avg_Close_Rate	-0.1890
Avg_Num_Products_Per_Opp	-0.0200
Avg_Term_Length	-0.1190
Total_Emails_Per_Week	-0.0923

Total_Prospect_Calls_Per_Week	No Correlation run because visuals did not support
Total_Customer_Calls_Per_Week	No Correlation run because visuals did not support
Total_Conversations_Per_Week	-0.0312
Total_Set_Demos_Per_Week	-01830
Total_Performed_Demos_Week	-0.1400
Total_Demo_Requests_Per_Week	-0.0300
Total_Online_Meetings_Per_Week	-0.2310

Model Identification and Changes

The project intention was to deploy a Supervised Machine Learning (Multiple Linear Regression) technique since labeled input and output training data was going to be used. Originally, I thought that I could use Python with the Scikit-learn package to perform a multiple linear regression (MLR).

- **Exception** - Since Databricks is an analytics platform on top of Apache Spark, I had to pivot and build regression models using the Spark packages and coding protocols. In addition to multiple linear regression, logistic regression, linear regression, and random forest techniques were applied to generate the best models / results.

Summary: Model Evaluation and Validation Process

I ran the project in two different Experiments: A and B for comparison purposes.

- **Experiment A** included preliminary runs of all the aggregated data across all sales divisions, using three different regression /validation combination models:
    - Logistic Regression (80/20 train and test split)
    - Logistic Regression (K-Fold validation)
    - Random Forest (K-Fold validation).

  Based on the performance of these models, feature engineering was applied to improve results in Experiment B.

- **Experiment B** included the feature engineering modifications, where I divided the data by Sales Division. New visualizations were created, and Pearson Correlation coefficients were performed on most correlated features. Experiment B consisted of:
    - Linear Regression for SMB (K-fold validation, t, and p testing).
    - Linear Regression for Mid-Market (K-fold validation t and p testing)
    - Multiple Linear Regression models for SMB (K-fold validation, t, and p testing)
    - Multiple Linear Regression models for MM (K-fold validation, t, and p testing)

Preliminary Results – Experiment A

Dependent Variable

The dependent variable “Percent To Quota” was used to leverage the vast 10-year company history of sales rep performance, including current as well as past employees. For all employees, the “Percent To Quota” was averaged throughout their tenure with the organization. It was important to include historical sales rep data because after the data cleaning process, since the sample size decreased from an n > 600, to n = 100.

It should be noted that “Percent To Quota” may appear overall lower than anticipated due to the hire and termination dates of sales representatives, which will affect their lifetime quota attainment percentage. Additionally, at the time of this report, there are 2 months of data that are excluded in the quota attainment percentage calculations which will also lower the values of the dependent variable. However, this will not alter the predictors and overall goal of the project and it is a reasonable standardized metric to use across all sales divisions with differing quotas.

In the future, I recommend another experiment using Total_Opportunity_Amount, by year, by sales rep, for further analytics because sales quotas were not added into the Salesforce data until more recently. More data could be leveraged, thereby increasing sample size, by eliminating the quota metric and focusing on revenue by division. This was not fully understood until results were analyzed, and time constraints prevented further exploration of this option.

Logistic Regression
Since the Average Opportunity Amount was correlated with higher Percent To Quota, I first ran a logistic regression to predict whether a sales representative was likely to achieve quota based solely on the average size of their opportunity.

- **Exception** – -The logistic regression is an exception to the original goal; however, since the initial correlations did not yield high level correlations with multiple features to run on MLR, I opted to use logistic regression.

Logistic Regression question: Assuming approximately 20% of the sales representatives make or exceed quota (top tier rep), can this predictive model determine which reps will perform in the top tier, based on the average size of their opportunities?

Summary of Logistic Regression Results:

- Sample size: n = 100
- Dependent variable: “Percent To Quota”
- Features: “Average Opportunity Amount”
- Validation:
  - Using 80/20 split: Area Under the Curve: AUC-ROC: 0.5378787878787878
  - Using K-Fold Validation: Area Under ROC: 0.6127185051235685

Experiment A1. Validation Data using the 80/20 split:

```
-------------+----------+--------------------+
|binary_column|prediction|         probability|
+-------------+----------+--------------------+
|            0|       0.0|[0.89202019038568...|
|            0|       0.0|[0.88875193924323...|
|            0|       0.0|[0.90217654990058...|
|            1|       0.0|[0.74485294797696...|
|            0|       0.0|[0.92871772742438...|
|            0|       1.0|[0.41813225403385...|
|            0|       0.0|[0.86650645249379...|
|            0|       0.0|[0.77892157094789...|
|            1|       0.0|[0.89476841645668...|
|            0|       0.0|[0.80553329544579...|
|            0|       0.0|[0.90272026823606...|
|            0|       0.0|[0.90568770261065...|
|            1|       1.0|[0.02175811099257...|
|            0|       0.0|[0.91258704206363...|
|            1|       0.0|[0.90548588680768...|
|            1|       0.0|[0.91134871020557...|
|            1|       0.0|[0.56228259329323...|
+-------------+----------+--------------------+
```

Logistic Regression Result Interpretation
Using the 80/20 split, the ROC-AUC of 0.53 is close to a random performance. The K-fold validation technique yielded better result of 0.6127185051235685, however, both results indicate that the model's ability to distinguish between the positive and negative classes is not much better than random chance in the logistic regression model. Since the logistic

regression technique did not yield significant results, I explored the possibility of other non-linear relationship that may exist among the variables.

Random Forest
I used Random Forest next since it tends to be robust to overfitting and is a better tool for complex relationships between the dependent variable and the features (Vidyha 2023).

Summary of Random Forest Results:

Sample size: n = 100
Dependent variable: "Percent To Quota"
Features: "Average Opportunity Amount," "Avg_Cust_Disc" and "Avg_Sell_Cycle"
Using K-Fold Validation:

- Root Mean Squared Error (RMSE): 0.12396054358188907
- Feature Importances:
- Avg_Opp_Amount: .5582428319309904
- Avg_Sell_Cycle: 0.23379201565840949
- Avg_Cust_Disc: 0.2079651524106001

Random Forest Result Interpretation
The "Average Opportunity Amount" was the most strongly correlated feature and the RMSE value was 0.12396. The RMSE provides an estimation of the model's performance in predicting he target value (SAP 2023), and normally, a value this low would be considered favorably. However, in the context of this project and the scale of the dependent variable (0-1O), I determined that the RMSE was too high to continue with any further validation on this model. In other words, the RMSE value represents on average, this model's predictions differ by about 12.4% of the range.

From this preliminary experiment, I explored options (Ray 2023) to improve its performance include the following:

- **Data Quality** – given the unanticipated small dataset (n = 100) which I had not anticipated, this affects the data. Also, data quality is dependent upon the reliance of sales reps and management adherence to data entry processes and requirements. Sales reps do not all record data in the same manner, and this affects the quality.

- **Feature Engineering** – I will explore additional relevant features that might contribute more information to the model. This entailed dividing the data up into sales division segments, which was part of the original proposal, however, with the n size of being so small, it could also affect results.

- **Alternative algorithms** – exploration of linear regression, multiple linear regression and random forest could help identify better model.

- **Validation** – K-fold validations are better used on small sample sizes; therefore, I will run this validation technique in subsequent runs.

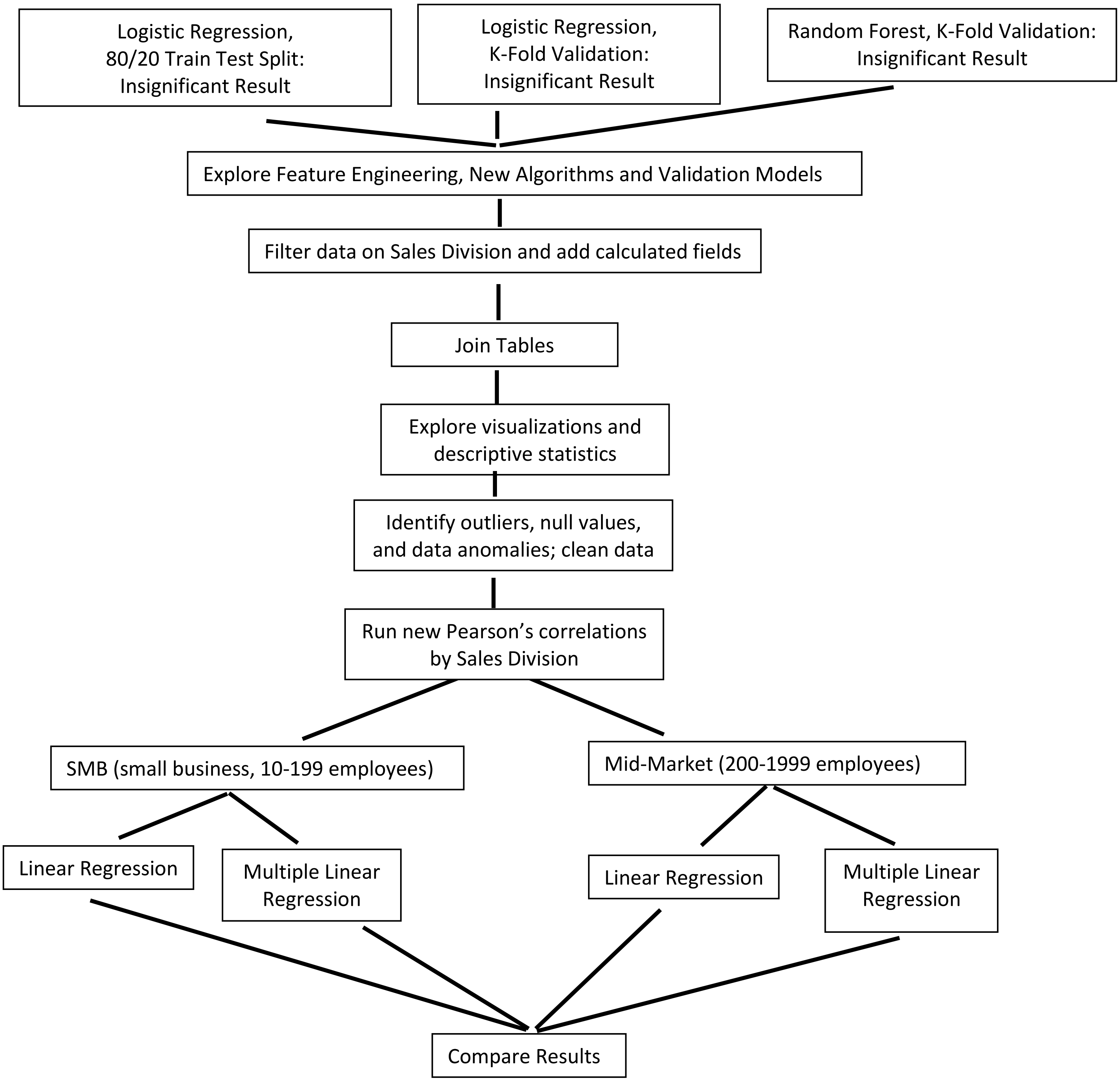
- **Hyperparameter values** – I could experiment with different hyperparameter values in the logistic regression model, but because the results were so poor, I did not do use this option.

- **Collect More Data** – this is not possible for 2023, given all available data was utilized for analysis. However, this project can be revisited in the future.

Next Steps: Experiment B

Since the results were insignificant, feature engineering and additional algorithms were explored, using K-fold validation to identify potential models. The workflow process employed is diagramed below.

Figure 1. Augmented Hypothesis Testing Process



Logistic Regression, 80/20 Train Test Split: Insignificant Result
Logistic Regression, K-Fold Validation: Insignificant Result
Random Forest, K-Fold Validation: Insignificant Result
Explore Feature Engineering, New Algorithms and Validation Models
Filter data on Sales Division and add calculated fields
Join Tables
Explore visualizations and descriptive statistics
Identify outliers, null values, and data anomalies; clean data
Run new Pearson’s correlations by Sales Division
SMB (small business, 10-199 employees)
Mid-Market (200-1999 employees)
Linear Regression
Multiple Linear Regression
Linear Regression
Multiple Linear Regression
Compare Results

Experiment B: Summary Changes

I made the following changes in Experiment B to improve correlation results.

- Segmented data by Sales Division
- Added Total Number of Accounts as possible predictor.
- Compared/Contrasted Sales Divisions as possible predictor.
- Removed outliers, null or anomalies contributing to possible noise.
- Compared new descriptive statistics and correlations.
- Re-ran algorithms based on highly correlated features in Pearson's correlation.
- Added P-Test and T-Test to the results, along with RSME.

Experiments B Correlations

Based on a set of new visualizations for each sales division, I ran the top Pearson's correlations once again which yielded different results from Experiment A (see below).

Experiment B. SMB Pearson's Correlation Results:

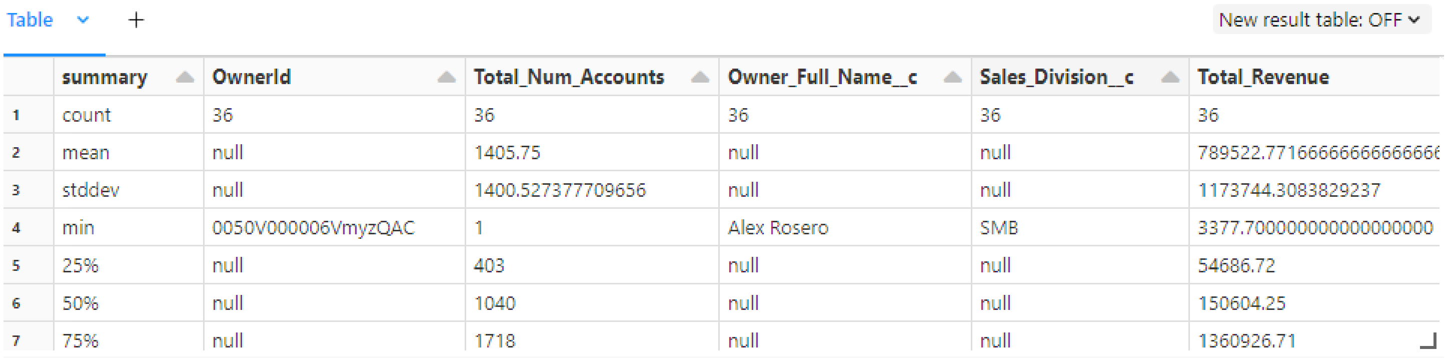
- Correlation between Percent To Quota and Average Opp Amount: 0.37218218071266135
- Correlation between Percent To quota and Average Cust Discount: 0.423386881349258
- Correlation between Percent To quota and Total Avg Weekly Activities: -0.20538575028381872

Experiment B. MM Pearson's Correlation Results:

- Correlation between Percent To Quota and Average Opp Amount: -0.12022508786589424
- Correlation between Percent To quota and Average Cust Discount: 0.31834755561553063
- Correlation between Percent To quota and Total Avg Weekly Conversations: 0.010416490833213958
- Correlation between Percent To quota and Total Number of Accounts: -0.27548335889903575
- Correlation between Percent To quota and Total Online Meetings Per Week: 0.2249661298099106

Descriptive Statistics

New descriptive statistics were run for both SMB and Mid-Market Sales divisions. A snapshot of 1/3 of the descriptive statistics table for SMB is pasted below. The complete descriptive statistics for both divisions are available in the Appendix.



	summary	OwnerId	Total_Num_Accounts	Owner_Full_Name__c	Sales_Division__c	Total_Revenue
1	count	36	36	36	36	36
2	mean	null	1405.75	null	null	789522.7716666666666666
3	stddev	null	1400.527377709656	null	null	1173744.3083829237
4	min	0050V000006VmyzQAC	1	Alex Rosero	SMB	3377.700000000000000000
5	25%	null	403	null	null	54686.72
6	50%	null	1040	null	null	150604.25
7	75%	null	1718	null	null	1360926.71

SMB Sales Division

The top two correlated features, based on the Pearson's correlation results, were identified as Average Opportunity Amount and Average Customer Discount. Regression Experiments will be set up to identify the best models using:

1. Experiment B1: Linear Regression – Average Customer Discount on Percent to Quota
2. Experiment B2: Multiple Linear Regression – Average Customer Discount and Average Opportunity Amount on Percent to Quota.

Mid-Market Sales Division
Regarding the Mid-Market segment, the “Average Cust Discount” and “Total Online Meetings per week” were selected as the top two predictors. Regression experiments will be set up to identify the best models using:

1. Experiment B3: Average Customer Discount on Percent to Quota
2. Experiment B4: Average Customer Discount and Total Online Meetings Per Week on Percent to Quota

The reason I did not select the negatively correlated, but significant feature “Number of Accounts,” is because the data had anomalies that unfortunately, could not be appropriately addressed in this project. For example, when a sales rep is terminated, territories of terminated reps are reassigned to others, creating a continually moving variable for “Number of Accounts” by rep . I therefore determined that this account data, by rep, may be skewed and unreliable for the purposes of this project’s analysis.

A similar issue was discovered with the once promising feature “Total Demo Requests Per Week.” Intuitively, one would think that the more demo requests (or leads) a rep is provided through the marketing department, the better their performance. However, this metric was also unreliable because of the vetting process of these leads. Often the website leads were produced in error, because of an unintentional click by the prospect. Also, leads are often assigned to the wrong sales representative and must be re-routed; yet the lead source and original lead owner may not necessarily change in the database. When this data is not corrected, it can be misleading, which is why I could not use this metric.

Table 2. Summary Results using “Percent to Quota” as the Dependent Variable

Division	Regression	Validation	Features	Results	T-Test and P-Test Results
All	Logistic	80/20	Avg_Opp_Amount	ROC = 0.5378	N/A = Did not perform because ROC was too low and comparable to random chance
All	Logistic	K-Fold	Avg_Opp_Amount	ROC = 0.6127	N/A – Did not perform because although improved, ROC value was still too low and comparable to random chance
All	Random Forest	K-Fold	Avg_Opp_Amount Avg_Sell_Cycle Avg_Cust_Disc	Root Mean Squared Error (RMSE): 0.12396054358188907 Feature Importances: Avg_Opp_Amount: .5582428319309904 Avg_Sell_Cycle: 0.23379201565840949 Avg_Cust_Disc: 0.2079651524106001	N/A – Did not perform because RMSE was too high given the Percent To Quota scale of 0-1
SMB	Linear	K-Fold	Avg_Cust_Disc	Root Mean Squared Error (RMSE): 0.11367827283577361 R-squared (R2): -0.939068077221632	Intercept: 0.21334034795651072 Coefficient 0: 0.00328007258819299T- Coefficient 0 : T-statistic = 2.5941296695425082, P-value = 0.014919053689924855 Coefficient 1: T-statistic = 6.560752017245371, P-value = 4.103771282792934e-07
SMB	Multiple	K-Fold	Avg_Opp_Amount Avg_Cust_Disc	Root Mean Squared Error (RMSE): 0.08129527906544977 R-squared (R2): 0.5695292561191588	Intercept: 0.16387927163782137 Coefficient 0: Estimate = 4.152605107643032e-06, T-value = 2.451509636119704, P-value = 0.020981244195025672 Coefficient 1: Estimate = 0.0032705427993116173, T-value = 2.810376139203201, P-value = 0.009096325199308364
MM	Linear	K-Fold	Avg_Cust_Disc	Root Mean Squared Error (RMSE): 0.13518621444979825 R-squared (R2): -0.07320713457562689	Intercept: 0.3226327201988182 Coefficient 0: T-statistic = 1.0496002531095305, P-value = 0.3105194110454699 Coefficient 1: T-statistic = 9.730423949339313, P-value = 7.150270620037702e-08
MM	Multiple	K-Fold	Avg_Cust_Disc Total_Online_Meetings_Per_Week	Root Mean Squared Error (RMSE): 0.1272495380644176 R-squared (R2): 0.04910805985992739	Intercept: 0.3123193943052424 Coefficient 0: Estimate = 0.3123193943052424, T-value = 0.22607205349133713, P-value = 0.8356722409559736 Coefficient 1: Estimate = 0.0006449522076622271, T-value = 0.20832566399276212, P-value = 0.8483167022891438

Results and Discussion

SMB Sales Division

Linear Regression Model

The Pearson's Correlation of the Average Customer Discount with Percent to Quota was the highest correlated feature at 0.423386881349258. The linear regression model's p-value was also significant (P-value = 4.103771282792934e-07).

The issue is when looking at the Root Mean Squared Error (RMSE) calculation (RMSE = 0.11367827283577361) and R-squared (R2) calculations (R2 = -0.939068077221632). RMSE or Root Mean Square Error measures the average error of the model's predictions. $R^{2,}$ or the "goodness of fit," is the proportion of the variation in the dependent variable ("Percent To Quota") that is predictable from the independent variable, or "Average Customer Discount." (Wikipedia 2023). R2 measures how well the regression predictions approximate the actual data points. The negative R2 value could indicate a poor fit of the model to the data. Additionally, the RMSE score is as high as that rejected in Experiment A when comparing it in context to a dependent variable scale of 0 to 1.

Conclusion on SMB Linear Regression Model

In conclusion, based on the combined results of R2 and RMSE despite the significance of the p-values, this model is rejected. Further exploration on the small sample size, outliers, non-linear relationships, or missing information that could affect the model's performance is warranted but is outside of the project's scope due to time constraints.

Given these results, I will accept the null hypotheses and conclude that Average Customer Discount is not correlated with a Sales Representative's quota attainment.

Multiple Regression Model

Using the two highest correlated features from the Pearson's correlation results (Average Opportunity Amount and Average Customer Discount), on Percent To Quota, the multiple linear regression model's p-value was statistically significant for both features (P-value for Avg_Opp_Amount = 0.020981244195025672; P-value for Avg_Cust_Disc = 0.009096325199308364), both of which are below the .05 minimum threshold for statistically significant findings. These results indicate that both variables are important predictors in the model.

Additionally, the Root Mean Squared Error (RMSE) result of 0.08129527906544977 is the lowest error of all models run, despite being higher than desired. The R-squared (R2) value of 0.5695292561191588 is also the best value obtained from all models run.

Given these results, I will reject the null hypotheses and conclude that Average Opportunity Amount and Average Customer Discount are positively correlated with a Sales Representative's quota attainment.

Multiple Regression Equation:

$Y = \beta 0 + \beta 1 X1 + \beta 2 X2$, where $\beta 0$ is the intercept, $\beta 1$ is X1 coefficient, $\beta 2$ is the X2 coefficient.

"Percent_To_Quota" = 0.16387927163782137 + 4.152605107643032e-06 x (Feature_0) + 0.0032705427993116173 x (Feature_1),
where Feature 0 = Avg_Opp_Amount and Feature 1 = Avg_Cust_Disc.

Discussion and Recommendations for SMB Sales Division

Although the sample size was small after separating data by sales division (n = 46), the model was still statistically significant. For future model improvement, additional data should be gathered from upcoming years of performance metrics. With a larger population size, new features may also become relevant and can be explored.

From an implementation perspective, company-wide communication and implementation of these practices are not necessarily advisable. While it is intuitive that the larger the deal size, on average, the more successful one will be, the issue comes with the feature Average Customer Discount. In both the linear and multiple linear regression models, Average Customer Discount was positively correlated with higher quota attainment results. In other words, the more a sales representative discounts their opportunities, the more likely they are to achieve or surpass quota. This is an important finding for internal leadership in that discounting policies may be re-visited and expanded to provide more discount flexibility in warranted circumstances. However, it is not advisable to communicate this finding to all sales representatives because of how it negatively impacts their average opportunity amount, which is also a correlated feature.

Mid-Market Sales Division Results and Discussion

Linear Regression Model

While the Pearson's correlation showed Average Customer Discount as highest of all features, the Linear Regression correlating Average Customer Discount on Percent To Quota yielded significant findings (P-value = 7.150270620037702e-08). However, please note that the p-value for the intercept was > .05 (P-value = 0.3105194110454699) which warranted further investigation into potential model issues, and its validity.

Based on some research (Stack Exchange 2015 and Research Gate 2015), it appears that when the value of the intercept is not significant, it implies that the data does not provide convincing evidence that the intercept is different from zero. However, this does not necessarily make the model invalid. It might be reasonable to have a model without a significant intercept, especially if the zero point for the feature (independent variable) does not have a meaningful interpretation. This seems conceivable in that without any average discount per opportunity, one could not yield a meaningful "Percent To Quota."

Goodness of Fit (R2)

However, the other issue is with the goodness of fit. The R-squared (R2) value of -0.0732 for a linear regression model with a dependent variable on a scale of 0 to 1 could indicate a poor fit of the model to the data. This finding for the Mid-Market sales division is like the SMB sales division's R2 result. R-squared is a measure of how well the independent variables, or features, explain the variability in the dependent variable. The negative R2 value could indicate a poor fit of the model to the data (Wikipedia 2023). The R-squared value should be between 0 and 1, where 0 indicates that the model does not explain any of the variability, and 1 indicates a perfect fit.

Root Mean Squared Error (RMSE)

The RMSE value of 0.13518621444979825 is even higher than those values which were initially rejected in Experiment A. I cited that those values were too high of an error, in context with a dependent variable scale of 0 to 1. The RMSE value is unacceptable for this model.

Conclusion on Mid-Market Linear Regression Model

In conclusion, based on the combined results of R2, RMSE and the p-value of the intercept, this model is rejected. Further exploration on the small sample size, outliers, non-linear relationships, or missing information that could improve the model's performance is warranted.

Given these results, I will accept the null hypotheses and conclude that Average Customer Discount is not correlated with a Sales Representative's quota attainment.

Multiple Linear Regression Model

For the Multiple Linear Regression Model on the Mid-Market Sales Division, "Average Customer Discount" and "Total Online Meetings Per Week" were used as the features on the dependent variable, " Percent To Quota." While the Pearson's correlation showed Average Customer Discount and Total Online Meetings Per Week as the highest correlation of all features, the Multiple Linear Regression correlating them with "Percent To Quota" did not yield significant findings (P-values= 0.8356722409559736 and 0.8483167022891438 respectively).

Given these results, I will accept the null hypotheses and conclude that Average Customer Discount and Total Online Meetings are not correlated with a Sales Representative's quota attainment.

Discussion and Recommendations for Mid-Market Sales Division
The results were not statistically significant because of two main reasons including:

1. The dataset was too small after dividing it into sales divisions, eliminating nulls and extracting outliers (n = 22). This is because there are fewer sales representatives in these Mid-Market roles, compounded with the fact that it is a newer role in the system, compared to the SMB roles. Further analysis can be performed in subsequent years, once more data is available. Furthermore, while the Pearson's correlation results yielded mild correlations with "Average Customer Discount" and "Total Online Meetings Per Week," this could be improved with the addition of more data.

2. Other features, such as Total Number of Accounts and Total Demo Requests Per Week, are areas to re-visit in the future as possible predictors. The challenges with these features are operational in nature. Changes in the operational practices will need to be explored to capture higher quality data in these two areas. Such changes include maintaining territories for at least 1 year for each sales representative and improved vetting of demo requests before they are routed to sales representatives. These two changes could yield interesting results for future analysis.

User Interface and Integration

Originally, the project user interface was to be limited to graphical representation of predictors and the thresholds required for Sales Standards of Excellence. Separate visualizations will be provided for the two sales markets (Small and Mid-Size) as well as the two verticals (Post-acute and Health & Human Services).

Exception – Due to the smaller than anticipated sample sizes, it was not possible to separate the data by vertical and run regressions accordingly. Instead, graphical representations of the data are provided below and in the sales report for SMB Sales Division Linear Regression and Multiple Linear Regression Models.

For the Mid-Market Sales division, since the features were not correlated to the dependent variable in either linear or multiple linear regression models, benchmark metrics are provided below as well as in the sales report.

SMB Sales Division Benchmarking
Below are the graphs for the best SMB Multiple Linear Regression Model. In addition to the deliverable this functional regression model, the following key benchmarking data will be provided in a Sales report to assist management in establishing sales metrics standards of achievement in the table below.

Figure 2. Multilinear Regression Graphs for SMB Sales Division

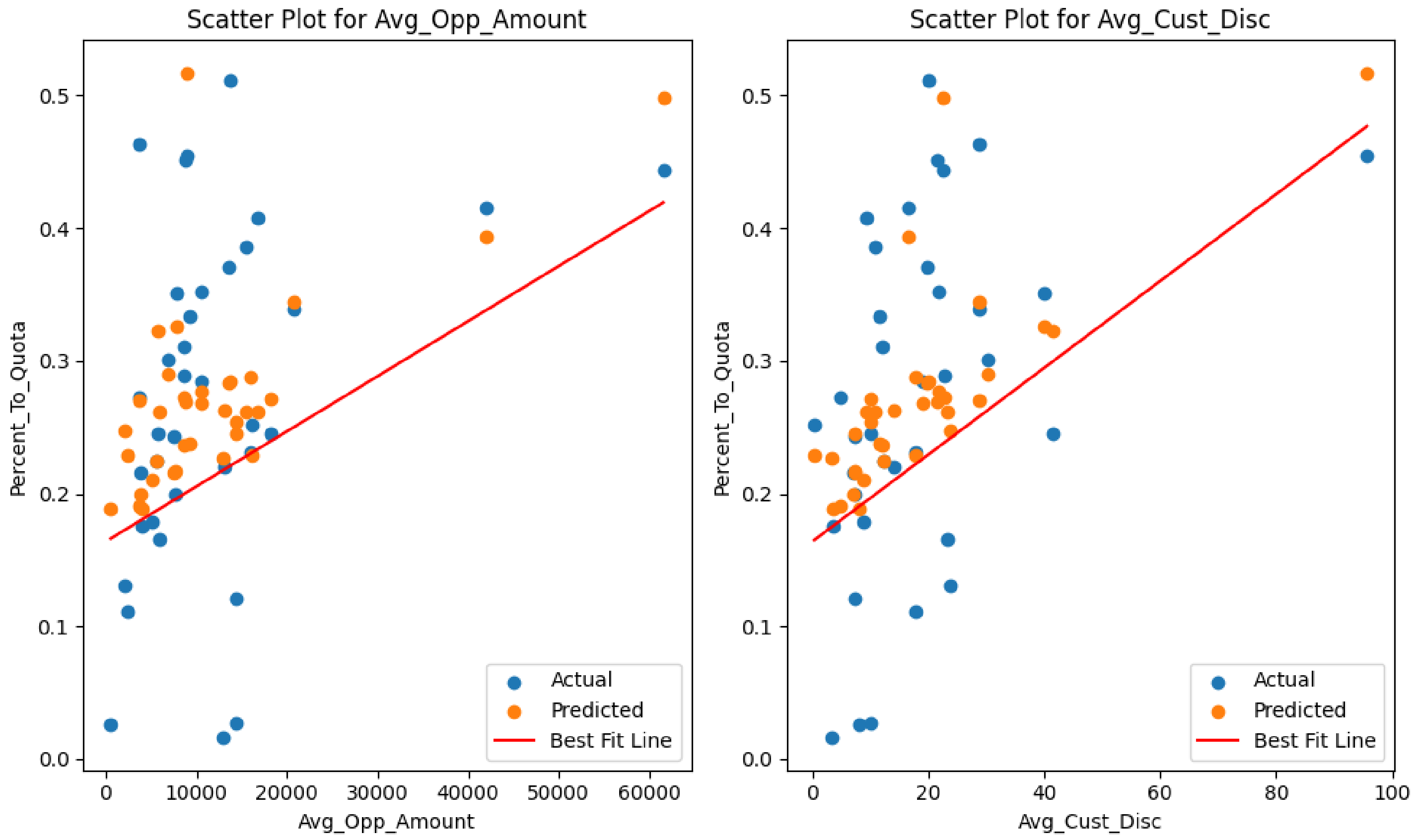


Table 2.SMB Summary Statics and Key Benchmarking Data

Summary Statistic	Average Opportunity Amount	Average Number of Products / Opp	Average Close Rate	Average Customer Discount	Total Activities / Week	Total Online Meetings/ Week	Total Demos Performed / Week
count	36	36	36	36	36		36
mean							
Std Deviation							
25th Percentile							
50th Percentile							
75th Percentile							

Mid-Market Sales Division Benchmarking

Although I was unable to create a significant regression model for the Mid-Market group, there is meaningful data available for benchmarking sales performance in key attribute areas. Key features selected for benchmarking are included in Table 3.

Table 3. Mid-Market Summary Statistics and Key Benchmarking Data.

Summary Statistic	Average Opportunity Amount	Average Number of Products / Opp	Average Close Rate	Average Customer Discount	Total Activities / Week	Total Online Meetings/Week	Total Demos Performed / Week
count	22	22	22	22	22	22	22
mean							
Std Deviation							
25th Percentile							
50th Percentile							
75th Percentile							

Integration Strategy

The user interface summarized above has been incorporated within a PowerPoint (identified as "SSE Sales Report Summary" with project results; this will be shared with Sales Leadership for their use as they deem appropriate. Sales leadership may use this data to:

- Adopt and roll-out new weekly activity metrics for sales representatives which are managed by our sales managers, using these new predictors and their corresponding thresholds determined by this machine learning project.

  Exception – Despite a working model for the SMB team, based on the results and smaller than anticipated sample sizes, I recommend using the descriptive statistics data as benchmarks for sales productivity.

- Re-align sales territories, if applicable, to provide each sales representative with ample sales opportunity, or "greenspace," accounts, states etc. to support future success.

  Exception – this is not recommended since "Accounts" could not be integrated as a viable feature in these models. This is because the number of accounts is fluid throughout the year, and not necessarily fixed (due to reassignments of accounts, territories etc.)

- Explore the possible integration of these new predictors into our weekly Sales Operations Management reports.

  Exception – While the SMB Sales Division does have a working regression model, its predictors may be controversial. It is recommended that sales leadership acknowledge the results and determine whether more discounting flexibility is warranted on larger deals within the SMB Sales division. At the same time, too much discounting adversely affects the Average Opportunity Amount or deal size. These two metrics should be monitored using the descriptive statistics as internal performance benchmarks only.

- Integrate sales management coaching in alignment with the project predictors and the threshold metrics required for success.

  Additional Recommendation - The benchmarking data is an incredibly good coaching tool in providing sales representatives with their performance metrics relative to their peers. I recommend using the 75-th percentile metrics as a goal for sales performance targets. This would provide sales with more context on what the top performers are achieving. Of course, each metric must be evaluated separately to determine its utility at the 75th percentile level.

- Depending upon the outcome, this “Sales Standards of Excellence” project could also be applied to our Account Management Team, as a separate but logical, compatible project.

  Additional Recommendation – Descriptive statistics were run on Account Management Sales Division data which will be provided to sales leadership for benchmarking purposes using the 75-th percentile metrics as a goal for sales performance targets.

Capstone Complexity

The original scope of this project was determined to meet or surpass the master’s level complexity requirements due to a variety of factors, including the use of a new Azure Databricks cloud-based environment (vs. desktop applications), mining of big data, over forty tables to analyze, data wrangling, joining of multiple tables, the machine learning application using multiple linear regression and finally, the potential significance of the project’s findings for our sales leadership. The SQL and Python notebook as well as integrated visualization software are all located within the Databricks platform.

Complexity Insight and Post-Project Perspective

This project was far more complex than originally detailed and meets or exceed the master’s level complexity requirements due to the additional factors beyond the original project’s scope:

- 43 SFDC object tables were analyzed.
- Nine tables were originally joined at the start of the project, using SQL. This strategy was ineffective due to the data consolidations and calculated field requirements warranted prior to joining the tables.
- Six tables were joined and used as the final project analysis due to the anomalies found in two tables which could not be effectively leveraged. Python was used to join these consolidated and cleaned data frames.
- Multiple calculated fields (30+), data formats, and general data wrangling required approximately 80% of the project’s time.
- Pyspark, rather than Python and Pandas, was used to prepare and clean the data before running regression experiments. PySpark is more suitable for complex data processing tasks that involve multiple stages of data transformation and analysis. Given the millions of rows of task data which required greater computing power, available with PySpark, this was the more appropriate software.
- PySpark’s libraries / packages, such as the linear regression and vector assembler packages, were also used in the regression analysis, rather than Python’s Scikit-Learn.
- Summary statistics were performed on all data to understand the tightness, distribution, and anomalies of the data.
- The PySpark and Databricks API alone added an unanticipated learning curve for this project, adding to its complexity. PySpark is complimentary to Python coding, but it is different, and adjustment were made to use the organizations big data computing capabilities, SFDC tables and overall capabilities.
- Seven different regression models were run on the data, which is far beyond what had originally been scoped. Originally, multiple linear regression models were identified as the most promising models, however, the following regression and validation models were performed to find the best model:

  1. Logistic regression – All Sales Divisions
     a. 80/20 training/test split validation
     b. R2, RMSE
  2. Logistic Regression – All Sales Divisions
     a. K-fold validations
     b. R2, RMSE
  3. Random Forest – All Sales Divisions
     a. K-fold validations
     b. R2, RMSE

4. Linear Regression: SMB
    a. K-fold validations
    b. R2, RMSE
    c. T and P-tests
5. Multiple Linear Regression: SMB
    a. K-fold validations
    b. R2, RMSE
    c. T and P-tests
6. Linear Regression: MM
    a. K-fold validations
    b. R2, RMSE
    c. T and P-tests
7. Multiple Linear Regression: MM
    a. K-fold validations
    b. R2, RMSE
    c. T and P-tests

- The data itself was extraordinarily complex in that there were millions of rows of activity data that had to be properly identified, labeled, consolidated, cleaned, and aggregated. Additionally, data results that were not logical (for example, the negative correlation of leads with Quota Attainment) had to be further analyzed and understood before abandoning the data as a viable predictor.
- The results interpretation was more complex than that learned in the master's program since it involved the significance of negative R2 values, the assessment of RMSE values that should be rejected based on dependent variable scale, and the understanding and implications of high intercept p values.

Appendix

Account Fields

Id
IsDeleted
MasterRecordId
Name
Type
RecordTypeId
ParentId
BillingStreet
BillingCity
BillingState
BillingPostalCode
BillingCountry
BillingLatitude
BillingLongitude
BillingGeocodeAccuracy
Phone
Fax
AccountNumber
Website
NumberOfEmployees
OwnerId
Days_Since_Last_Account_Activity__c

contact
LastName
FirstName

Forecastingquota
Id
QuotaOwnerId
PeriodId
StartDate
ProductFamily
CurrencyIsoCode
QuotaAmount
QuotaQuantity
QuotaOwnerId
IsQuantity
IsAmount
CreatedDate
CreatedById
LastModifiedDate
LastModifiedB
SystemModstamp
ForecastingTypeId
Territory2Id

Opportunity
Id
IsDeleted
AccountId
RecordTypeId
IsPrivate
Name
Description
StageName
Amount
Probability
ExpectedRevenue
TotalOpportunityQuantity
CloseDate
Type
NextStep
LeadSource
IsClosed
IsWon
ForecastCategory
ForecastCategoryName
Deal_Term_Months__c
Number_of_Opportunity_Products__c
Partner_Discount__c
Owner_Full_Name__c
License_Total__c

Pricebook2Id
OwnerId
Territory2Id
IsExcludedFromTerritory2Filter
CreatedDate
CreatedById
LastModifiedDate
LastModifiedById
SystemModstamp
LastActivityDate
PushCount
LastStageChangeDate
FiscalQuarter
FiscalYear
Fiscal
ContactId
LastViewedDate
LastReferencedDate

ConnectionReceivedId
ConnectionSentId
ContractId
HasOpenActivity
HasOverdueTask
LastAm

Type_of_Opportunity__c
Expected_Close_Date__c
Extended_Term_Total__c
License_Total__c
Subscription_Total__c
Intacct_Entity__c
IntacctID__c
Vertical__c
Account_Type__c
Contract_Start_Date__c
Total_Days_Open__c
Age__c
Owner_Full_Name__c
Weighted_Amount__c
Billing_State_Province__c
Number_of_Opportunity_Products__c
Deal_Term_Months__c

sbqq__quote__c
SBQQ__AverageCustomerDiscount__c

task
# need type, name, related to, subject; keep all fields and do not delete
#use task assigned, subject = demo request alert for demo requests metrics
Id
RecordTypeId
WhoId
WhatId
Subject
ActivityDate , also run min, max, days_employed, type/day
Status = Completed
Priority
IsHighPriority
OwnerId
Description
Type
IsDeleted
AccountId
IsClosed

opportunitylineitem
OpportunityId
Billing_Frequency__c
Vertical__c

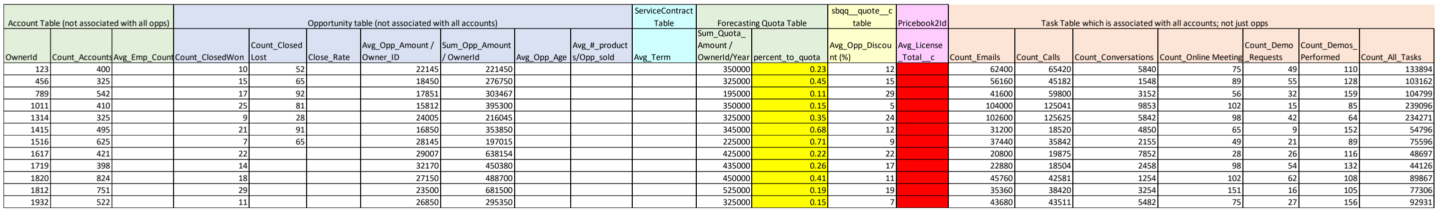
servicecontract
Quote_Price_Cap__c
AccountId
Term

rubybenchmarking
Account_Name_Client_Size (#enterprise, mid-market etc.)

Account_Owner (#name of rep)
Account_Name_Vertical
Account_Name_18_digit_ID (key)

Opportunityhistory
Id
OpportunityId
CloseDate

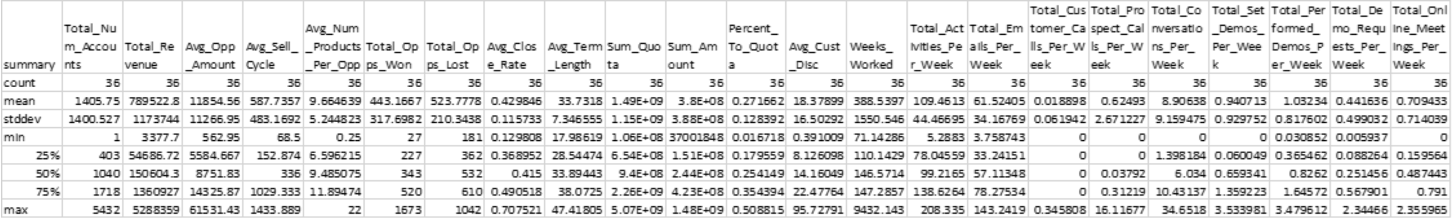
Diagram 1 – mockup visual of seven tables with desired final data frame metrics



Account Table (not associated with all opps)			Opportunity table (not associated with all accounts)							ServiceContract Table	Forecasting Quota Table		sbqq__quote__c table	Pricebook2Id	Task Table which is associated with all accounts; not just opps						
OwnerId	Count_Accounts	Avg_Emp_Count	Count_ClosedWon	Count_Closed Lost	Close_Rate	Avg_Opp_Amount / Owner_ID	Sum_Opp_Amount / OwnerId	Avg_Opp_Age	Avg_#_products/Opp_sold	Avg_Term	Sum_Quota_ Amount / OwnerId/Year	percent_to_quota	Avg_Opp_Discount (%)	Avg_License_Total__c	Count_Emails	Count_Calls	Count_Conversations	Count_Online Meeting	Count_Demo_Requests	Count_Demos_Performed	Count_All_Tasks
123	400		10	52		22145	221450				350000	0.23	12	🟥	62400	65420	5840	75	49	110	133894
456	325		15	65		18450	276750				325000	0.45	15	🟥	56160	45182	1548	89	55	128	103162
789	542		17	92		17851	303467				195000	0.11	29	🟥	41600	59800	3152	56	32	159	104799
1011	410		25	81		15812	395300				350000	0.15	5	🟥	104000	125041	9853	102	15	85	239096
1314	325		9	28		24005	216045				325000	0.35	24	🟥	102600	125625	5842	98	42	64	234271
1415	495		21	91		16850	353850				345000	0.68	12	🟥	31200	18520	4850	65	9	152	54796
1516	625		7	65		28145	197015				225000	0.71	9	🟥	37440	35842	2155	49	21	89	75596
1617	421		22			29007	638154				425000	0.22	22	🟥	20800	19875	7852	28	26	116	48697
1719	398		14			32170	450380				435000	0.26	17	🟥	22880	18504	2458	98	54	132	44126
1820	824		18			27150	488700				450000	0.41	11	🟥	45760	42581	1254	102	62	108	89867
1812	751		29			23500	681500				525000	0.19	19	🟥	35360	38420	3254	151	16	105	77306
1932	522		11			26850	295350				325000	0.15	7	🟥	43680	43511	5482	75	27	156	92931
]

Descriptive Statistics

Table 3. Descriptive Statistics for Experiment B – SMB Sales Division



summary	Total_Num_Accounts	Total_Revenue	Avg_Opp_Amount	Avg_Sell_Cycle	Avg_Num_Products_Per_Opp	Total_Opps_Won	Total_Opps_Lost	Avg_Close_Rate	Avg_Term_Length	Sum_Quota	Sum_Amount	Percent_To_Quota	Avg_Cust_Disc	Weeks_Worked	Total_Activities_Per_Week	Total_Emails_Per_Week	Total_Customer_Calls_Per_Week	Total_Prospect_Calls_Per_Week	Total_Conversations_Per_Week	Total_Set_Demos_Per_Week	Total_Performed_Demos_Per_Week	Total_Demo_Requests_Per_Week	Total_Online_Meetings_Per_Week
count	36	36	36	36	36	36	36	36	36	36	36	36	36	36	36	36	36	36	36	36	36	36	36
mean	1405.75	789522.8	11854.56	587.7357	9.664639	443.1667	523.7778	0.429846	33.7318	1.49E+09	3.8E+08	0.271662	18.37899	388.5397	109.4613	61.52405	0.018898	0.62493	8.90638	0.940713	1.03234	0.441636	0.709433
stddev	1400.527	1173744	11266.95	483.1692	5.244823	317.6982	210.3438	0.115733	7.346555	1.15E+09	3.88E+08	0.128392	16.50292	1550.546	44.46695	34.16769	0.061942	2.671227	9.159475	0.929752	0.817602	0.499032	0.714039
min	1	3377.7	562.95	68.5	0.25	27	181	0.129808	17.98619	1.06E+08	37001848	0.016718	0.391009	71.14286	5.2883	3.758743	0	0	0	0	0.030852	0.005937	0
25%	403	54686.72	5584.667	152.874	6.596215	227	362	0.368952	28.54474	6.54E+08	1.51E+08	0.179559	8.126098	110.1429	78.04559	33.24151	0	0	1.398184	0.060049	0.365462	0.088264	0.159564
50%	1040	150604.3	8751.83	336	9.485075	343	532	0.415	33.89443	9.4E+08	2.44E+08	0.254149	14.16049	146.5714	99.2165	57.11348	0	0.03792	6.034	0.659341	0.8262	0.251456	0.487443
75%	1718	1360927	14325.87	1029.333	11.89474	520	610	0.490518	38.0725	2.26E+09	4.23E+08	0.354394	22.47764	147.2857	138.6264	78.27534	0	0.31219	10.43137	1.359223	1.64572	0.567901	0.791
max	5432	5288359	61531.43	1433.889	22	1673	1042	0.707521	47.41805	5.07E+09	1.48E+09	0.508815	95.72791	9432.143	208.335	143.2419	0.345808	16.11677	34.6518	3.533981	3.479612	2.34466	2.355965

Table 4. Descriptive Statistics for Experiment B – MM Sales Division

summary	Total_Num_Accounts	Owner_Full_Name__c	Sales_Division__c	Total_Revenue	Avg_Opp_Amount	Avg_Sell_Cycle	Avg_Num_Products_Per_Opp	Total_Opps_Won	Total_Opps_Lost	Avg_Close_Rate	Avg_Term_Length	Sum_Quota	Sum_Amount	Percent_To_Quota	Avg_Cust_Disc	Weeks_Worked	Total_Activities_Per_Week	Total_Emails_Per_Week	Total_Customer_Calls_Per_Week	Total_Prospect_Calls_Per_Week	Total_Conversations_Per_Week	Total_Set_Demos_Per_Week	Total_Performed_Demos_Per_Week	Total_Demo_Requests_Per_Week	Total_Online_Meetings_Per_Week
count	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22
mean	1084.455	null	null	1752095	28156.34	944.896	12.15246	499.7727	565.2727	0.440551	36.65357	1.57E+09	4.96E+08	0.330748	21.35331	552.026	112.1941	66.73564	0.009951	0.156028	10.326	0.785048	0.863651	0.354867	0.64314
stddev	904.3352	null	null	5684056	16331.04	504.2868	5.787109	367.3984	217.6249	0.101498	5.414118	1.25E+09	4.35E+08	0.10617	18.21201	1983.6	46.81782	33.16598	0.026489	0.243223	8.934884	0.842456	0.8077	0.391134	0.71861
min	19	Alicia Renc	Mid-Marke	8149.73	8149.73	118.7039	3.640261	144	258	0.181595	28.45822	3.1E+08	1.09E+08	0.130604	4.898037	71.14286	5.2883	3.758743	0	0	0	0	0.030852	0.005937	0
25%	471	null	null	43534.25	17760.42	533.1429	6	308	375	0.370569	33.87805	6.84E+08	2.28E+08	0.231488	11.68766	104.2857	94.00776	47.30941	0	0	4.599034	0.060049	0.365462	0.108632	0.129001
50%	894	null	null	131602.5	21933.75	1191.25	12.42857	366	554	0.421137	35.83924	9.4E+08	2.86E+08	0.332903	20.0265	147.1429	99.2165	61.48953	0	0.033882	9.079612	0.448109	0.552941	0.21978	0.421359
75%	1398	null	null	341890.5	39193.74	1365	18	526	667	0.491098	40.54902	2.16E+09	6.62E+08	0.415753	23.19539	147.2857	142.4739	78.31481	0	0.278371	13.22222	1.269641	1.31165	0.457115	0.741176
max	4168	Will Sapp	Mid-Marke	26357669	73008.17	1499.5	21	1673	1042	0.677712	47.41805	5.07E+09	1.48E+09	0.508815	95.72791	9432.143	208.335	143.2419	0.105882	0.977692	34.6518	3.214563	3.479612	1.748266	2.355965

References

Mckinney, Wes (2022). Python for Data Analysis, Third Edition. Published by O'Reilly Media,

Geron, Aurelien (2023). Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, 3rd Edition. Published by O'Reilly Media Inc.

Lubanovic, Bill (2020). Introducing Python, Second Edition. Published by O'Reilly Media, Inc.

VanderPlas, Jake. (December 2016). Python Data Science Handbook, First Edition. Published by O'Reilly Media, Inc.

Matthes, Eric (2019). Python Crash Course, Second Edition. Published by No Starch Press.

.

Etaati, Leila (June 26, 2018). Azure Databricks Part 2. Published by Radacad.com.

Sruthi E R (Updated October 26, 2023). Understand Random Forest Algorithms With Examples. Analytics Vidhya (https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=A.%20Random%20Forest%20is%20a,and%20outliers%20in%20the%20data.)

Almaliki, Zaid Alisse (March 19, 2019). Do you know how to choose the right learning algorithm array among 7 different types? Towards Data Science (www.towardsdatascience.com)

VanDerwerf, Paul (August 26, 2020). Train-Test Split for Evaluating Machine Learning Algorithms. Machine Learning Mastery (www.machinelearningmastery.com)

Sap.com (2023). SAP Predictive Analytics https://help.sap.com/docs/SAP_PREDICTIVE_ANALYTICS/41d1a6d4e7574e32b815f1cc87c00f42/5e5198fd4afe4ae5b48fefe0d3161810.html

Ray, Sunil (Updated September 25, 2023). 8 Ways to Improve Accuracy of Machine Learning Models. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/#:~:text=There%20are%20several%20ways%20to,bagging%2C%20boosting%2C%20and%20stacking.

Stackexchange (2015). Linear regression: Intercept isn't significant. https://stats.stackexchange.com/questions/160628/linear-regression-intercept-isnt-significant

Wikipedia (2023). Coefficient of Determination.https://en.wikipedia.org/wiki/Coefficient_of_determination#:~:text=There%20are%20cases%20where%20R,fitting%20procedure%20using%20those%20data.

Databricks.com

Sprark.apache.org

Coding Assistance and Troubleshooting:
Databricks Assistant (integrated AI assistant in Databricks to troubleshoot coding errors and issues)
Google.com
Stackoverflow.com