Capstone 2: Mainline Information Systems Sales Forecast Project

Who? What? When? How Much?

Sales forecasting is one of the most vital components of a well run sales program and one of the hardest parts of the sales manager's job. In fact it's listed as the second most difficult part of a sales manager's job with the most difficult being motivating the sales team.

Successful sales managers understand which deals their reps are working on and use information about those deals to project what business will close. Timely, accurate predictions help companies with goal-setting, hiring, budgeting, and other activities that affect cash flow and the bottom line. The impact of poor projections bears negative financial consequences, particularly for publicly traded companies, since stock values rise and and fall in correlation with hit or miss scenarios.

Background

Mainline Information Systems, Inc. (MIS) is a technology VAR (value-added reseller) serving thousands of enterprise and mid-market U.S. clients. Customers trust in Mainline's 30-plus years of experience in providing best-fit technology solutions, whether deployed on site or in the cloud. Mainline is a privately held corporation headquartered in Tallahassee, Florida.

Sales and C-suite executives within Mainline engage in typical, continuous forecasting scenarios based on sales pipeline data. Management is looking for a tool that predicts future sales based on information about deals in the pipeline, saving the company both time and money.

Sales opportunities are entered into Salesforce.com, a commonly used CRM platform among companies worldwide. Information about the customer is also stored in Salesforce, including industry, number of employees, Mainline's salesperson assigned to the customer, the sales territory the salesperson/customer is in, and history on all opportunities for that customer.

Opportunities data includes the opportunity name, which is a concatenation of customer name and a brief description of the product or solution to be sold. It also includes the potential sales close date, whether the opportunity has been billed and when, the revenue amount, and a calculated field on the number of days between creation date and anticipated close date.

This project will **use feature analysis on historical data to identify a model for predicting the completion of sales in the current pipeline.** The data set prior to cleaning is more than 97,000 rows and 20 columns, representing opportunities created

in Salesforce for the past five years.

A Closer Look at the Data

After dropping redundant columns, deleting rows or columns not relevant to the project, formatting, cleaning and adding new feature columns (shown in blue), the resulting data set includes more than 82,000 rows and 18 columns:

Categorical Data

Column Name	Column Description
Account Name	name of the customer
Opportunity Name	concatenation of Account Name with brief description of deal
SuccessfulSale	(target) Boolean; 1 if the deal closed, 0 otherwise
Has Products	Boolean; 1 if the deal has products loaded , 0 otherwise
Region	the sales territory to which the account and Opportunity Owner/salesperson belong
Industry	industry in which the customer operates Opportunity
Owner	the salesperson on the opportunity/deal
100 days 200 days 300 days 400 days 500 days >500 days	Boolean; bins created to represent whether the opportunity reached the "age" described; 1 for true, 0 for false; age is defined as the number of days between the creation date of the opportunity and the anticipated close date (for example, an opportunity that is 400 days old has 1's in the 100 days, 200 days, 300 days and 400 days columns and 0s in the 500 days and >500 days columns)

Continuous Data

Column Name	Column Description
Employees	the customer's employee count; representative of size
Amount	the estimated revenue on the deal
Perc_Billed	the # of closed(won) deals divided by total deals for the account
Perc_AE_Billed	the # of closed (won) deals divided by total deals for salesperson
Rank	assigned value to an Account based on $ Amount of closed deals in data set; top customer has rank of 1 and so on

Target Variable: SuccessfulSale

The data set includes a higher proportion of won deals (73%) than lost deals (27%), which seems atypical for a sales company. Since this data set reflects 5 years of deals tracked in Salesforce, it might be assumed that deals with low potential for closing may have never been loaded.

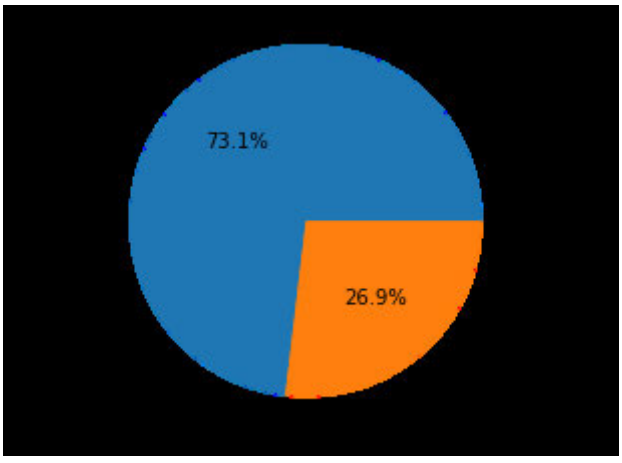


Figure 1: Percentage of Sales Won to Sales Lost in Data Set

Identifying Relationship of Other Variables to Target Variable

Since the goal is to predict whether an opportunity will result in a won deal, correlations between the other columns, called features, and the target column are explored.

Correlation is a statistical measure, ranging from negative 1 to positive 1, that expresses the extent to which two variables are linearly related. Values close to zero represent a weak relationship. Values closer to positive 1 or negative 1 describe direct or inverse relationships, respectively.

Correlations

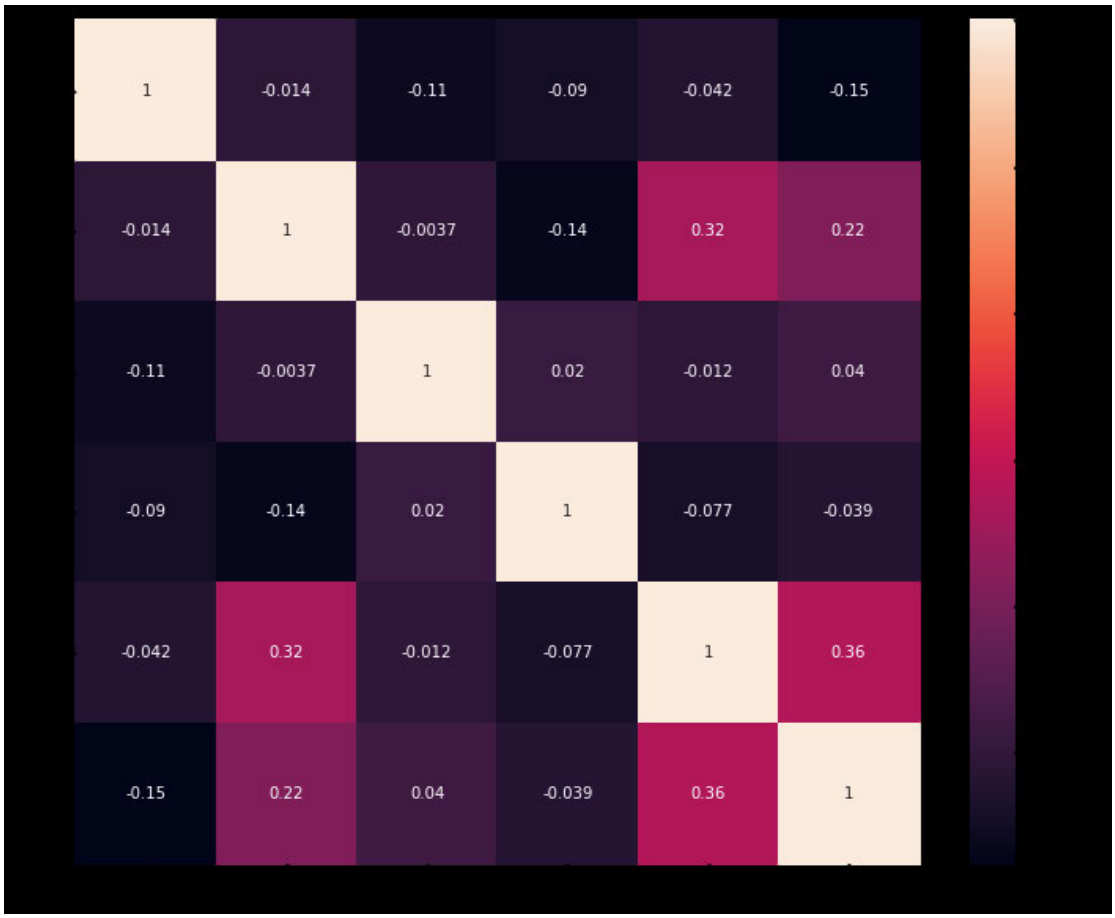
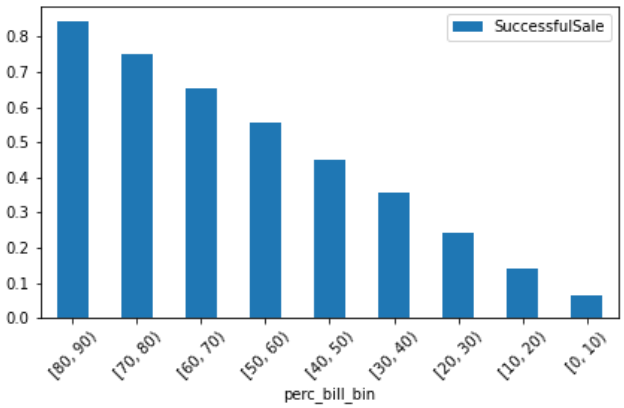


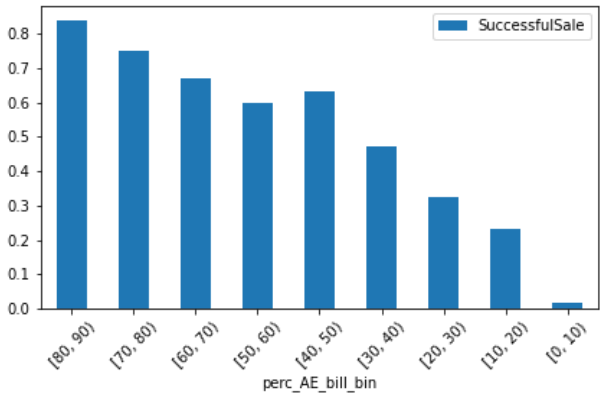
Figure 2. Correlation Heatmap for Continuous Variables

Figure 3. Correlation Heatmap for Categorical Variables

Close Ratios (Perc_Bill, Perc_AE_Bill)

The Perc_Billed and Perc_AE_Billed, which can be referred to as close ratios for Accounts and Opportunity Owners, respectively, have meaningful positive correlations with a successful sale. For any given opportunity with an unknown outcome, it is more likely that a deal will close if it is with a customer with a high proportion of won deals within the data set. The same is true based on the success rate of the deal's Opportunity Owner (salesperson). Since an Account generally has the same salesperson, although not always the case, these features are interrelated.





Figures 4 and 5: Percent of Won Deals by Close Ratio for Account, Account Executive

Further statistical testing on these two variables with SuccessfulSale data indicate significant relationships that would be unlikely just by chance.

In the categorical matrix, the Region and Industry to which a sales opportunity belongs are not key drivers of the target variable.

Days (Age of Opportunity)

The second matrix on categorical features shows correlations between the days categories and SucccessfulSale, the target variable. Note that the number of days an opportunity remains in the sales pipeline shows a weaker relationship with the successful close of a deal as the number of days increases. Age has statistical significance in relation to success.

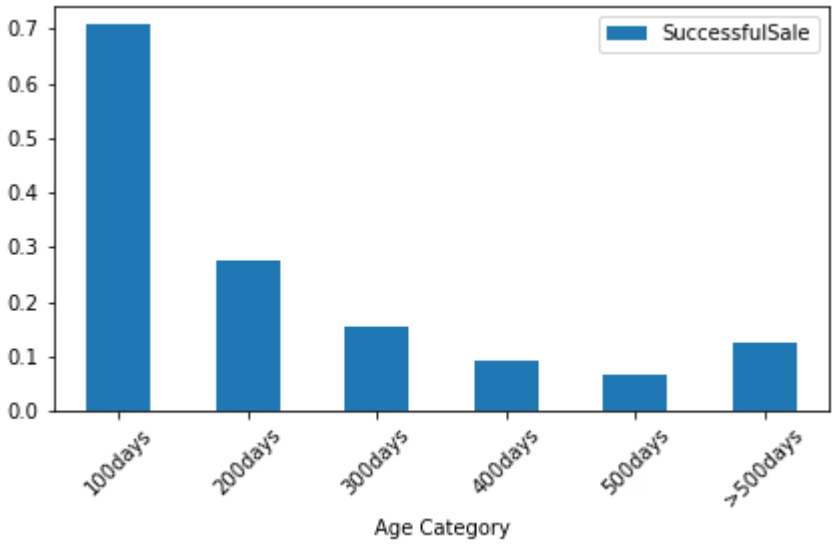


Figure 6: Percent of Won Deals by Age Category

Has Products

Another feature that shows a correlation as strong as Perc_AE_Billed is Has Products. When this Boolean value is zero, there is virtually no chance of a won deal whereas there's a 70% chance of a successful sale when products are loaded to the opportunity.

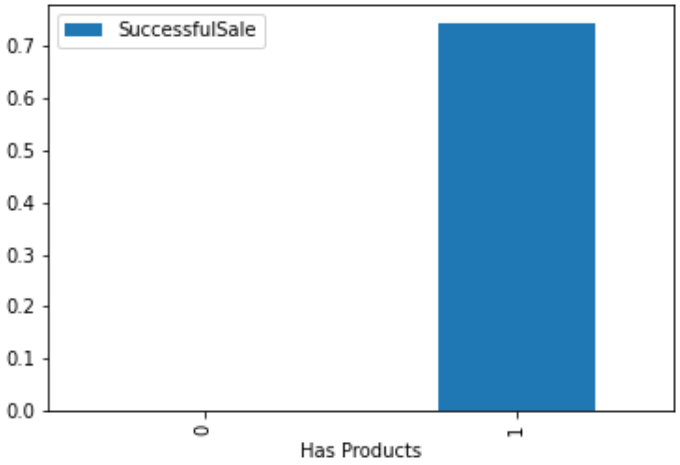


Figure 7: Percent Won Deals with Products

A Deeper Look at Amount

Since Mainline deals can be very large, it seems that the size of a deal might be important to a successful sale. Breaking up the Amount column, which has opportunities as large as $50 million, may be helpful for understanding the typical size of won deals.

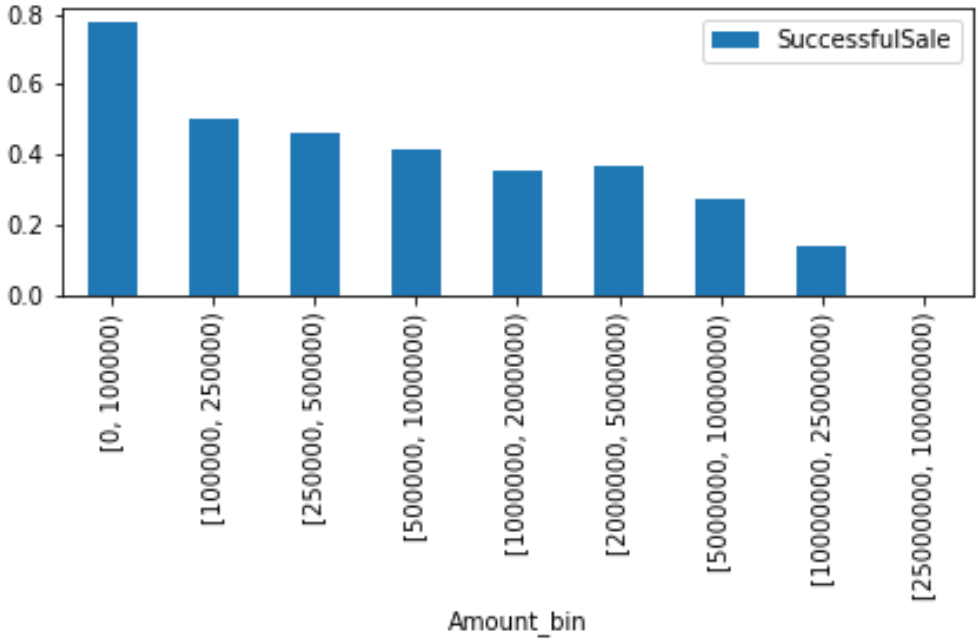


Figure 8: Percent Won by Amounts

The data suggest that smaller deals have a higher success rate. Those under $100k close nearly 80% of the time while those greater than $100k have a 50-50 chance of closing. This most likely represents the most common deal size, which forms the “bread and butter” of Mainline’s business.

Which Features Matter Most?

Correlation figures provide clues as to which features should be explored further. Since relationships exist between both continuous and categorical values, data science tools appropriate for handling both data types are necessary. Two widely used resources for classification problems are the Random Forest and Gradient Boosting models, which use **decision trees** on different samples.

The main differences between these classifiers is the order in which the trees are built and the way the results are combined. Random Forest classifiers build independent trees and then average their majority votes. Gradient Boosting classifiers build and assess performance one tree at a time, each new tree correcting errors made by the previously trained tree.

These tools include a helpful assessment on which features in a dataset are most important to predicting the assigned target variable. Running each on the Mainline data set reveals the most important features for predicting whether a deal will close are the following:

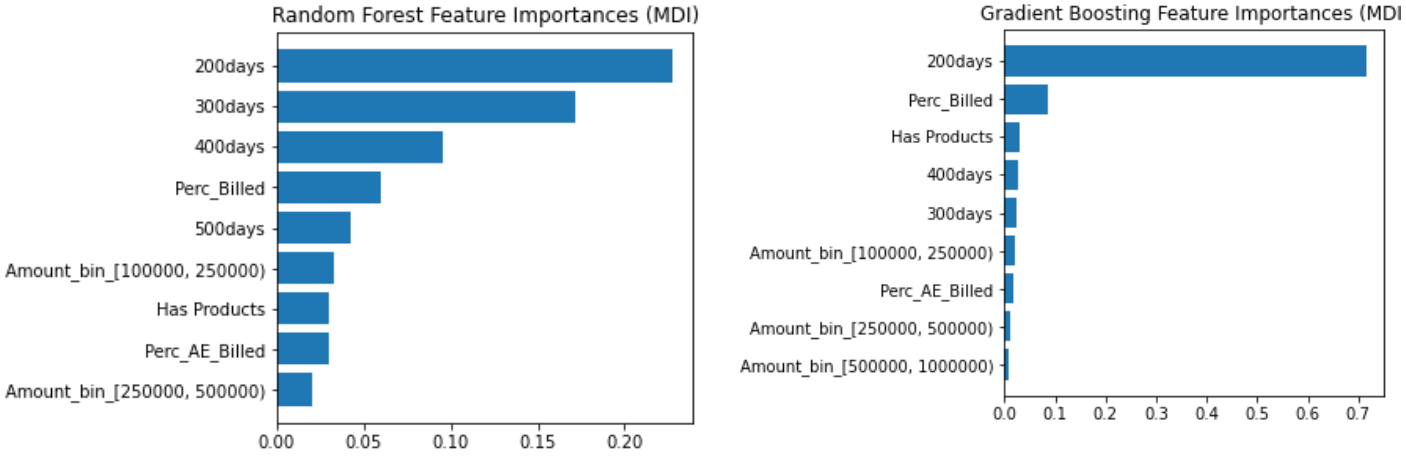


Figure 9: Most Important Features for Predicting Opportunity Outcome

It is reassuring to see the most important features are the same as identified in the heatmaps showing correlations between features and the target variable, SuccessfulSale. Note that the two algorithms weigh or rank the features differently based on how each produces and assesses the outcomes of its decision trees.

The age and amount of an opportunity (measured in days categories and amount categories, respectively), the close ratios (Perc_Billed for Account, Perc_AE_Billed for Opportunity Owner) and whether the opportunity Has Products loaded to it are the most important features. According to the Gradient Boosting classifier, whether an opportunity will close depends heavily on that opportunity being no more than 200 days old.

A Note on Age

Choosing to include age invites a small amount of data leakage into these and other classifying models used in this project. For each opportunity in the training data, when it closed or didn't close and at what age is known. However, these models will be run on data that hasn't yet reached its final age, and so when it is determined not to have an age of 500 days, it is unknown whether or not it will reach that age eventually. As such, one would expect the probability of a win to decrease as an opportunity gets older, which seems intuitive. It also seems helpful to know what the age is in terms of predicting, versus the alternative of simply not using the age at all.

A choice is made here to train the models on data that is different from what they would be predicting on. To clarify, the models have knowledge about the true final age of an opportunity that they wouldn't normally have when predicting on unseen data. In the future, it might make sense to try comparing these models to ones that don't use age at all on real world data.

Models: Which One?

The next stage involves identifying an appropriate model that will perform the best at predicting “Won” deals. Models designed for binary classification problems are required for the Mainline data.

Model performance can be evaluated a number of ways, but the most appropriate measurement in this case is the Receiver Operating Characteristic curve, or ROC curve. It is a plot of the false positive rate versus the true positive rate. The area under the plotted curve, referred to as AUC, can be used to measure a model’s performance. The higher the AUC, the better the model’s performance at distinguishing between two classes (i.e. “Won” or “Lost”). If a model has a score of 1, it can perfectly classify all the data.

One thing to consider is the imbalanced nature of the Mainline data. Seventy-three percent of the opportunities are won deals. A model may perform better if it has equal representation of both outcomes. This can be achieved by producing an equal number of samples from each class for model testing. Another group of samples that are representative of the ratio of won deals to lost deals can also be used.

Each group of samples is split into training and testing sets, then applied to the model. The training data is used to train the model and the test data is used to evaluate performance. Several models known as classifiers are tried and evaluated using AUC scores.

ROC AUC SCORES		
For Different Sampling Methods		
Model	Even	Representative
Random Forest	0.885	0.884
Logistic Regression	0.873	0.852
Ada Boost	0.894	0.889
Gradient Boosting	**0.901**	0.899

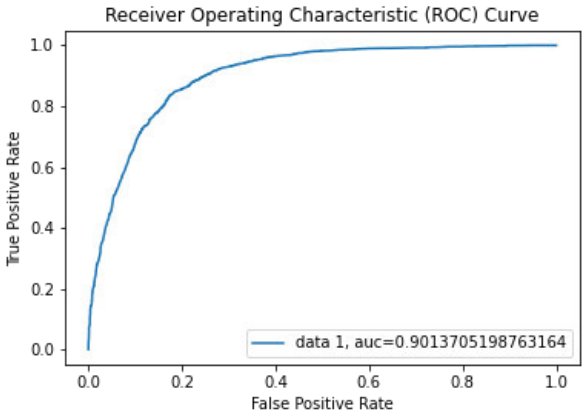


Figure 10: ROC AUC scores for different classifiers, samplings; ROC curve for Gradient Boosting

The best model in this case is the Gradient Boosting Classifier since it has the highest AUC score, although the others are not vastly different. Using an equal number of samples from each class also produces better results.

Improving Model's Performance

While an AUC score of 0.901 is great, it might be possible to improve this statistic by tuning the model's settings. Research indicates the most crucial parameters for decision tree models are the learning rate, the number of trees and tree depths.

Testing the model parameters one by one, the best performance occurs at a learning rate of 0.5, tree count of 200, and tree depth of 7. In combination, however, the learning rate should be adjusted in proportion to tree count by a ratio of 1:10.

Using GridSearch, multiple values for each parameter can be tested in combination to isolate a set of those values that produce the best score. In this case, a learning rate of 0.05, 1500 trees with a maximum depth of 7 produced the highest AUC score, 0.908.

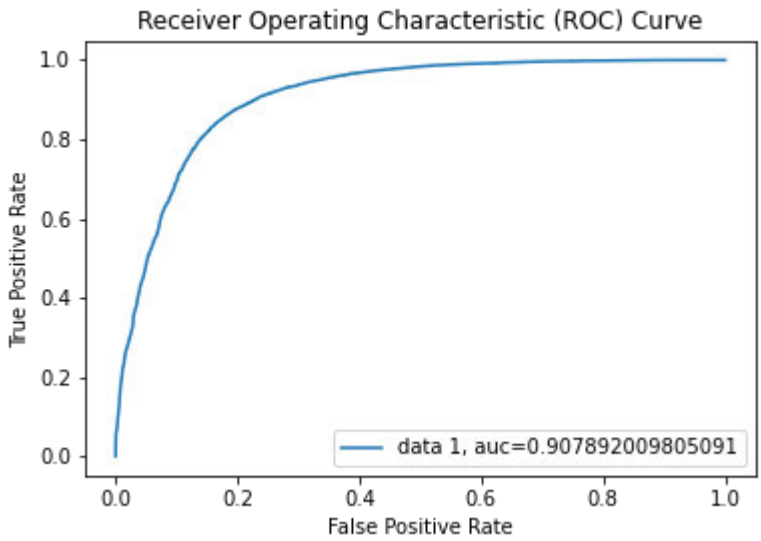


Figure 11: Results of Gradient Boosting Classifier with tuned parameters on full data set

Final Model, Predicting Outcomes

For predicting outcomes, it is often more appropriate to use the probabilities generated by the model and setting a threshold for classification. Using probabilities allows for control over the threshold for determining which outcome an opportunity will be assigned. For example, if an opportunity has a 75% probability of resulting in a sale, it can be classed to "Won." Otherwise, a model called to predict the outcomes will use a threshold of 50% to assign classes.

Testing different thresholds indicates the best classification results are when the threshold is set to the default of probability >= 0.5. A comparison between the actual and predicted outcomes shows that **the model correctly classifies 88% of the opportunity outcomes.**

Classification	Counts	% of Total
Correct	72,842	88.3%
Incorrect	9,670	11.7%
Grand Total	**82,512**	**100.0%**

	Predicted Lost	Predicted Won
Lost Oppty	**15,377**	6,797
Won Oppty	2,873	**57,459**

Won deals have very high probability values whereas lost deals have very low values. The large standard deviation for Lost opportunities indicates many of these might be incorrectly classified. In fact, 31% of Lost deals are misclassified as compared to only 5% for Won deals within the Mainline data set.

Type	# Correctly Classified	Probability % Mean	StDev
Lost Oppty	15,377	0.15	0.13
Won Oppty	57,459	0.90	0.09
Grand Total	**72,836**		

While the purpose of this project, to predict whether an opportunity will result in a sale, has been accomplished, C-suite executives, who frequently evaluate success with dollars, might want to know more, such as, what is the dollar value of the prediction?

So What: Show Me the $$$

The original MIS data set includes the Amount for each opportunity. Recall that this value is the estimated revenue for a deal. With the model's predictions, Revenue Forecasts may be calculated. A breakdown by year shows the following:

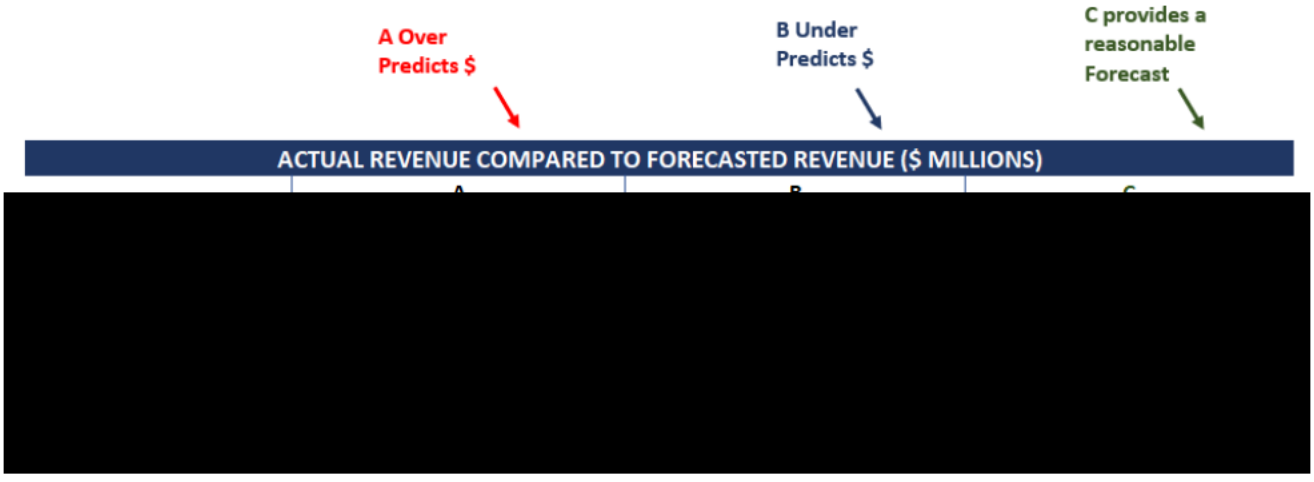


Figure 12: Revenue Forecasts based on predicted outcomes, probabilities

Note that the known outcomes of the data are disregarded here. Solely the predicted outcomes, whether correct or incorrect, are used to produce revenue forecasts. In the real world, only the probabilities generated by the model would be known. These probabilities prove useful for producing revenue forecasts that aren't over stated. Multiplying the Amounts by the assigned Boolean outcome values of 1 generates an aggressive forecast (A in figure above). However, multiplying the Amounts by the probabilities creates a more conservative result (B).

It is up to sales management to determine and report revenue forecasts. Most managers would agree that forecasting too high is undesirable. However, a habit of forecasting too low might imply excessive hedging. Adjustments to these calculated approaches might be used instead. In the example in Figure 12, using approach A as an upper bound and approach B as a lower bound, a manager might use approach C, the average of both, to make a better forecast without going over. Other methods might involve rounding the probabilities prior to calculating forecasted values.

Summary

Recall that the objective of this project was to identify which opportunities in a sales pipeline will result in won business. After cleaning and adding features to a large data set with known sales outcomes, several classifying models were trained and tested on a large sample of the data, then evaluated. The chosen model, a tuned Gradient Boosting Classifier, correctly predicted the outcomes for 88% of the MIS data.

Using the predicted probabilities for opportunities classified as "won" proved relevant for forecasting revenues, perhaps one of the most important uses for the modeled data. Further testing of the model on unseen data is necessary to determine whether it is suitable for production. Research indicates that sales managers spend 2.5 hours a week on forecasting activities. What if that could be accomplished in 15 minutes?

While not the scope of this project, the insights discovered herein might help with the tracking and interpretation of Mainline's sales pipeline and related forecasts. Features deemed important by the models' algorithms that aren't currently used for forecasting could be considered.

Reflections, Thank Yous

The field of data science is vast and growing. This project was done with just a few months of online training, which included learning Python, long nights, and sheer will to push through it. The experience made it clear that there is so much more to learn than what was required for this task.

It is likely that more knowledge was gained by tackling this work-related project than doing a project on a public data set. Public data sets are addressed by many data science enthusiasts, who document their findings online. Since Mainline data is private,

the work of others on the same data didn’t exist as a resource. This made the project more challenging and rewarding. The results are authentically original.

Great admiration and appreciation are extended to mentor, Ben Bell, whose guidance and support were crucial to the success of this project.The most helpful online references were articles on binary classification problems and imbalanced data. Stack Overflow was a primary resource for how-tos and trouble-shooting code errors. A great deal of support is available to anyone willing to reach out or search for it. Thank you to all students and professionals whose online postings prove incredibly valuable to others.