Boosting Sales by Pursuing Promising Opportunities

Alex Spence (ajs811), Jonas Bartl (jb6868), Xiao Li (xl998)
From Germany with Love, Center for Data Science, New York University
Advisor: Professor Brian Dalessandro, MBA

***Abstract*—A finely tuned gradient boosting model was developed in accordance with data science/machine learning industry standards to predict the top 500 sales opportunities available to an automotive supplier. The model was chosen because it showed the highest accuracy based on AUC score, and the highest profit based on profit curve methodology. It is recommended to deploy this model at a company and re-calibrate as needed based on the deployment. In the future, this model can be sold to companies in other industries who are looking to optimize their sales opportunities.**

I. Business Understanding

THE underlying business problem of the automotive supply industry is whether to pursue a sales opportunity, specifically a client who may be interested in purchasing automotive supplies. A wholesale automotive supplier must identify which opportunities to pursue such that after pursuit of all sales opportunities, the company profits. Thus, the importance of identifying which sales opportunities are likely to be won cannot be overstated as the company's performance in this area will ultimately decide whether the company will profit and thrive, or operate under a deficit and falter. Our team of data scientists developed a data mining solution to aide wholesale automotive suppliers in this critical aspect of the business: sales opportunity selection. A model was developed based on historical sales opportunity data, such as revenue from client, region, route to market, etc., to predict the probability that a potential sales opportunity will be won. This model can be used by a wholesale supplier to improve decision making on which sales to focus its efforts towards. This would lead to efficient resource usage and an increase in sales/profits.

For purpose this study, we will make the following assumptions:

- The likelihood of winning a sales opportunity is positively correlated with the sales staffs' effort (unobserved), meaning that focusing on designated opportunities and pursuing those with more effort will increase the probability of winning these opportunities.
- Pursuing a sales opportunity is always related with some cost (e.g. evaluation the customer's needs, preparing a tender) that would not occur otherwise.
- The sales department of the company is facing limited resources that do not allow the pursuit of all sales opportunities with maximum effort.

This makes it reasonable to pursue only a subset of all sales opportunities and putting more effort on those which we set to be 500. We will come back to different scenarios we analyzed and discuss the decision making rule which opportunities to pursue more in detail in section *Results and Model Selection*.

II. Data Understanding

Our data stems from a IBM's watson and is avaible online: https://www.ibm.com/communities/analytics/watson-analytics-blog/sales-win-loss-sample-dataset/. Originally, the data probably stems from a customer relationship management software, but we do not have detailed knowledge about this.

The target variable of our model is the opportunity result (whether the sales opportunity is won or lost). For each sales opportunity (the instances), the associated features of past sales opportunities are defined below:

1) Supplies Group – supplies group of the the supplies the client is interested in purchasing, including:
   - Car Accessories
   - Car Electronics
   - Performance and Non-auto
   - Tires and Wheels
2) Supplies Subgroup – the subgroup of the supplies the client is looking to purchase, including:
   - Batteries and Accessories
   - Car Electronics
   - Exterior Accessories
   - Garage and Car Care
   - Interior Accessories
   - Motorcycle Parts
   - Performance Parts
   - Replacement Parts
   - Shelters and RV
   - Tires and Wheels
   - Towing and Hitches
3) Region – the following locations of the sales opportunity (USA):
   - Mid-Atlantic
   - Midwest
   - Northeast
   - Northwest
   - Pacific
   - Southeast
   - Southwest
4) Route to Market – lists how the sales opportunity came to be known to the company:
   - Field Sales
   - Reseller
   - Telecoverage
   - Telesales
   - Other
5) Sales Stage Change Count – The number of times a sales opportunity changes stages.
6) Elapsed Days in Sales Stage – The number of days between the change in sales stages. The counter is reset for each new sales stage.
7) Total Days Identified Through Closing – Cumulative number of days the sales opportunity has spent in the following sales stages:
   - Identified/Validating
   - Gained Agreement/Closing
8) Days Identified Through Qualified – Cumulative number of days the sales opportunity has spent in the following sales stages:
   - Identified/Validating
   - Qualified/Gaining Agreement
9) Opportunity Amount (USD) - Cumulative sum of supply pricing that the client is interested in purchasing
10) Client Size by Revenue - The client's revenue separated into five categories:
    - 1 – under $1,000,000
    - 2 – $1,000,000 to $9,999,999.99
    - 3 – $10,000,000 to $49,999,999.99
    - 4 – $50,000,000 to $99,999,999.99
    - 5 – $100,000,000 and over
11) Client Size by Employee Count – The client's number of employees separated into five categories:
    - 1 – under 1,000 employees
    - 2 – 1,000 to 4,999 employees
    - 3 – 5,000 to 9,999 employees
    - 4 – 10,000 to 29,999 employees
    - 5 – 30,000 employees and over
12) Revenue from Client Past Two Years - The total sales which the wholesale automotive distributor has had with a particular client in the past two years separated into four categories:
    - 1 – under $50,000
    - 2 – $50,000 to $399,999.99
    - 3 – $400,000 to $1,499,999.99
    - 4 – $1,500,000 and over
13) Competitor Type – A binary variable which is "Known" and "Unknown".
14) Ratio Days Identified to Total Days – The number of days the sales opportunity has been identified divided

by the number of days the opportunity was in the sales pipeline.

15) Ratio Days Validated to Total Days – The number of days the sales opportunity was validated divided by the number of days the opportunity was in the sales pipeline-
16) Ratio Days Qualified to Total Days – The number of days the sales opportunity was qualified divided by the number of days the opportunity was in the sales pipeline.
17) Deal Size Category – A grouped version of the opportunity amount:
    - 1 – under $10,000
    - 2 – $10,000 to $24,999.99
    - 3 – $25,000 to $49,999.99
    - 4 – $50,000 to $99,999.99
    - 5 – $100,000 to $249,999.99
    - 6 – $250,000 to $499,999.99
    - 7 – $500,000 and over

There may be some upfront costs for a wholesale automobile supplier to gather the above historical data to train the model. If the above data has not been collected over the years, the company may have to do some research on past sales to get the relevant information. In addition, if the data has been collected, but is unorganized, the company may have to set up scripts to clean the raw data. The company must evaluate whether using the model is worth the cost of starting the historical data pipeline. However, once the scripts have been set up to collect historical data, the ongoing maintenance costs with processing future data should be minimal.

III. Exploratory Analysis

We will start our analysis by taking a look at the distribution of sales opportunities won and lost across different deal sizes. Figure 1 shows that the company loses more sale opportunities than it wins over all groups of in the deal sizes, reflecting a base rate of 23% overall won cases. Even worse, of all deals with a volume of $ 50,000 - $ 100,000 and $ 100,000 - $ 250,000 which make the largest groups in terms of their deal volume, the company wins the least opportunities in relative term (13% and 18%, respectively).

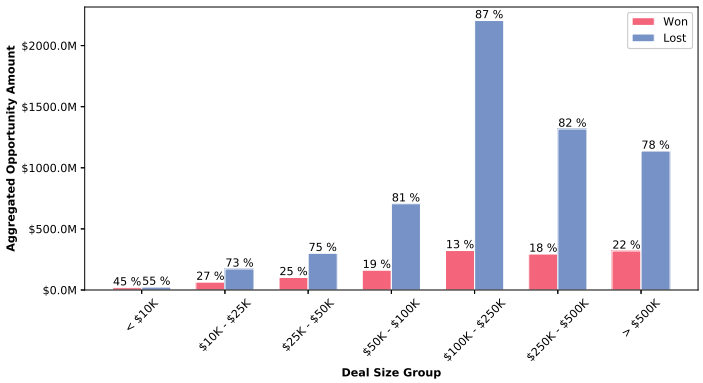

Fig. 1. Distribution of Won and Lost cases by Deal Size

Next, we want to explore the relevance of past revenues. Intuition suggests that on average it should be easier to win a tender with a client with whom a firm has made business in the past. Our dataset confirms this hypothesis with a win rate of 52.6 % in the sales opportunities with revenue in the past, compared to a winrate of only 12.4 % in sales opportunities where the company has no relationship with the customer. As a heuristic, this would suggest to focus mainly on those sales opportunities where there has been revenue in the past.

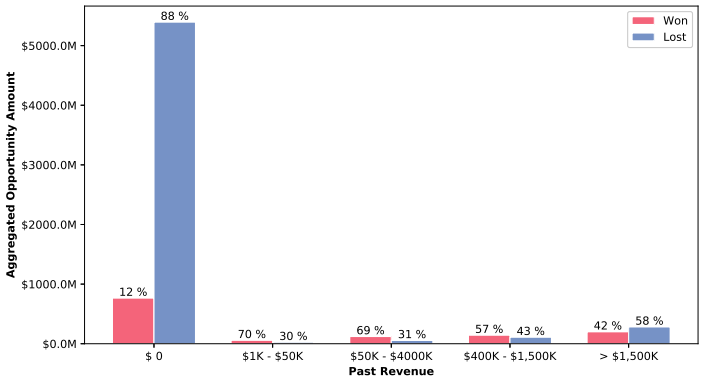

Fig. 2. Distribution of Won and Lost cases by Past Revenue

However, figure 2 shows that, by far the biggest sales volume lies in cases without past revenues and that even in those cases with past revenue, the share of won cases declines with the deal size. This is intuitive as competition

for large sales volumes may be higher than for lower sales volumes.

Based on these descriptive insights, we will build our predictive model as outlined in the next sections.

IV. Data Preparation

Although the raw data was relatively clean and organized, some data preparation was necessary to ready the data for model development. First, the target variable, "Opportunity Result" was modified such that "won" values had a value of "1" and "loss" values had a value of "0". Then, the dataset was checked for any null values, of which there none. However, there were 2047 instances with a opportunity amount of '0'. These instances were removed since it is not profitable to pursue opportunities which have no sales potential.[1] Since string values are not compatible with most model developments, categorical features were transformed to indicator variables.

After this, any features subject to leak were removed from the data set. This included: Sales Stage Change Count, Elapsed Days In Sales Stage, Ratio Days Identified To Total Days, Ratio Days Qualified To Total Days, Ratio Days Validated To Total Days, Total Days Identified Through Closing, and Total Days Identified Through Qualified. The model would be subject to leakage if these features were included because the values of these features are not known until the target variable is known.

Once this was completed, the dataset was normalized using the "preprocessing.StandardScaler()" method from the ScikitLearn library. Next, the data was randomly split into training (80%), validation (10%), and test (10%) sets to be used in model development. We considered using cross-validation techniques, but decided it was not necessary because our dataset has more than 30-50 data points per feature (each of our training/validation/test sets have at least 7,597 data points per feature), the base rate is more than 5% (our data set - 23%), and after feature selection the ratio of features to entries is less than 50.[2]

[1]In practice, it may beneficial to pursue tenders that do not lead to revenues but improve the brand awareness or the public image of a company. However, for the purpose of this study, we will ignore this possibility.

V. Modeling and Evaluation

Since our business problem is a class probability estimation, we chose to focus on some of the more common predictive modeling techniques used in the industry which support this type of problem: decision tree classifier, logistic regression, random forest, and gradient boosting. In addition, since we are attempting to predict the 500 top sales opportunities, this is a ranking problem. So in accordance with best practices, we decided to use Area Under the ROC Curve as our maximizing loss function.

Our strategy was first to compare baseline models based on maximizing the AUC on the validation set. Then, we used grid search to find the best parameters for each classifier, and used feature importance, and mutual information to perform feature reduction. The hyperparameters/feature set combination for each model was selected based on which combination produced the highest AUC score on the validation set. In addition, a profit curve was developed which was fitted on the training data and predicted on the validation data.

Choosing the best model was a two-part approach. First, the model with the highest AUC score on the validation set was selected. Then, the model with the highest profit corresponding to 500 sales opportunities was selected. The best model selected was the one with the highest AUC score and the highest profit potential for the top 500 sales opportunities. Once the model was selected, a new training set (training + validation sets) was created to fit to the selected model. Finally, we checked for generalization by using this model which was fitted on the new training set to predict the outcomes of the test set. A ROC curve with AUC score and a profit curve was created for this generalization measure.

[2]However, to prove this, we performed cross-validation with one of the baseline models to prove that the results would be relatively equal.

A. Feature Reduction

Two naive methods were used to select the most informative attributes: feature importance and mutual information. Scikit learn's DecisionTree fit function was used to determine the feature importance for each feature. Figure 3 shows the feature importances.

Based on these results, we selected the top 22 features according to their feature importance (features with feature importance > 0.01) which were used as a separate feature set in each modeling scenario.

Feature Importance by Decision Tree

	0
Opportunity Amount USD	0.3944503635862426
Revenue From Client Past Two Years	0.15678935263528876
Client Size By Employee Count	0.052173509882155224
Client Size By Revenue	0.050199088814370754
Route To Market Reseller indicator	0.021210876637279587
Region Midwest indicator	0.02052149127810987
Supplies Subgroup Exterior Accessories indicator	0.020123970202406773
Supplies Subgroup Garage & Car Care indicator	0.0182415187909204
Region Northwest indicator	0.017812099234304064
Region Pacific indicator	0.017688702539467457
Supplies Subgroup Batteries & Accessories indicator	0.01670934721519237
Region Southeast indicator	0.01667030672931753
Supplies Subgroup Replacement Parts indicator	0.016504885077680425
Supplies Subgroup Motorcycle Parts indicator	0.015822137036863564
Route To Market Fields Sales indicator	0.015236853062994746
Region Southwest indicator	0.015119945173431624
Competitor Type Unknown indicator	0.01448634551616171
Competitor Type None indicator	0.0139283049093987
Region Northeast indicator	0.013524091874027951
Region Mid-Atlantic indicator	0.012918000471953869
Supplies Subgroup Shelters & RV indicator	0.011433313735725084
Competitor Type Known indicator	0.011262513241058552
Supplies Subgroup Towing & Hitches indicator	0.00948569725179704
Supplies Subgroup Interior Accessories indicator	0.009381440272835684
Supplies Group Performance & Non-auto indicator	0.008712246688067459
Supplies Group Car Accessories indicator	0.007744650754854662
Route To Market Other indicator	0.0070524518398833325
Supplies Subgroup Performance Parts indicator	0.005983877469874108
Route To Market Telesales indicator	0.004644614714073093
Supplies Subgroup Car Electronics indicator	0.001078413501705578
Route To Market Telecoverage indicator	0.0009457176450778374
Supplies Group Tires & Wheels indicator	0.0008701520544238938
Supplies Group Car Electronics indicator	0.000773806367786362
Supplies Subgroup Tires & Wheels indicator	0.0004999137137282411

Fig. 3. Ranked Feature Importance

Similarly, mutual information was used to determine the amount of information on the target variable we could gain from observing a single feature.

Figure 4 shows the mutual information of each feature. It is interesting to see that the besides 'Opportunity Amount' and 'Revenue From Client Last Two Years', the other features' ranking in the two tables vary considerably. And if we choose features based on the same rule applied to feature importance (choose features that have importance > 0.01), then we end up with four features in total.

Mutual Information

	0
Opportunity Amount USD	0.12310365458024886
Revenue From Client Past Two Years	0.061697190995661844
Route To Market Reseller indicator	0.010731754004644989
Route To Market Fields Sales indicator	0.010215385229776608
Region Midwest indicator	0.008532667642922931
Supplies Subgroup Batteries & Accessories indicator	0.008486125190131366
Supplies Group Performance & Non-auto indicator	0.00769298462369461
Competitor Type None indicator	0.00685853865158248
Supplies Subgroup Shelters & RV indicator	0.006592540460844232
Route To Market Telecoverage indicator	0.006334532591943276
Competitor Type Unknown indicator	0.0059840184121664475
Supplies Subgroup Exterior Accessories indicator	0.0055945195681423865
Competitor Type Known indicator	0.005459714127229898
Supplies Group Car Accessories indicator	0.005268441110311439
Supplies Subgroup Replacement Parts indicator	0.005061326541268674
Supplies Subgroup Garage & Car Care indicator	0.004854478804840356
Client Size By Revenue	0.004741427867156478
Region Southeast indicator	0.004535018732050977
Supplies Subgroup Motorcycle Parts indicator	0.004325508934265532
Region Mid-Atlantic indicator	0.004110368027041789
Supplies Subgroup Interior Accessories indicator	0.003916809264273402
Region Northeast indicator	0.003914166407850805
Region Pacific indicator	0.0037301732930374865
Route To Market Telesales indicator	0.003192232613598911
Client Size By Employee Count	0.003169522793134094
Supplies Subgroup Performance Parts indicator	0.0022435264050997272
Supplies Subgroup Car Electronics indicator	0.002183876928301798
Supplies Subgroup Towing & Hitches indicator	0.0020302300263275175
Supplies Subgroup Tires & Wheels indicator	0.0018523481249825835
Region Southwest indicator	0.0017328292641369814
Supplies Group Tires & Wheels indicator	0.0013133576804302827
Route To Market Other indicator	0.0011090589286655295
Region Northwest indicator	0.0008095410271404813
Supplies Group Car Electronics indicator	0.0

Fig. 4. Ranked Mutual Information

B. Decision Tree Classifier

The decision tree classifier was chosen as a model to test since its binary nature in splitting on the most important features in accordance with maximizing information gain would be very intuitive in a business environment. In addition, it is easy to interpret, implement, and prediction/scoring is relatively inexpensive. One disadvantage is that decision tree classifiers can be easy to overfit. We avoided this by carefully tuning the parameters such that the predicted values and AUC scores were computed using the validation set instead of the training set.

C. Logistic Regression

It was decided to use logistic regression as a modeling technique because it is well suited for ranking and has strong statistical properties in that it is naturally defined to give a probability distribution.

D. Random Forest

The random forest classifier was selected as modeling technique for the same reasons as the decision tree classifier.

The random forest classifier develops many decision trees and leverages each to potentially develop better models. We thought it would do a decent job of ranking the top 500 sales opportunities.

E. Gradient Boosting

Similarly, the gradient boosting classifier was chosen for the same reasons as random forest and decision tree. We thought gradient boosting may give us better results than random forest since each decision tree it trains is successively used to train the next decision tree. The fact that gradient boosting is not as susceptible to overfitting is both a good news and bad news for us: the good in that we could simply set the number of boosting stage very large to obtaining better result; the bad in that parameter tuning for gradient boosting can be very slow since the more boosting stages there are, the slower the fit becomes.

F. Baseline Models

Figures 5, 6, and 7 show the ROC curves using default setting of each classifier before feature reduction, after feature reduction by feature importance, and after feature reduction by mutual information, respectively.

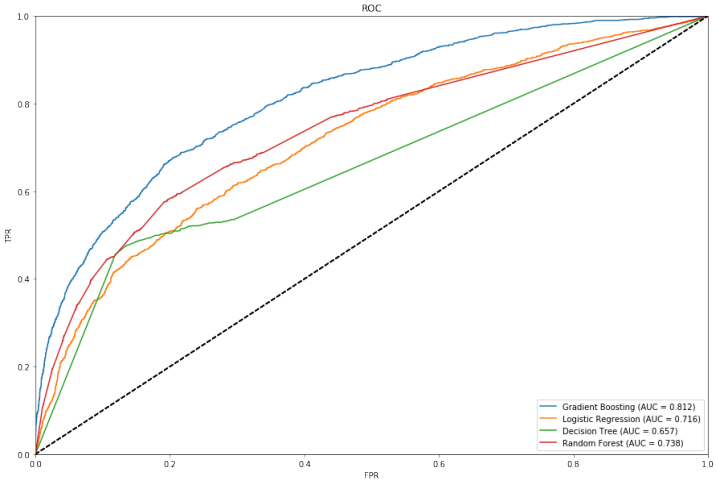

Fig. 5. Baseline model without feature reduction

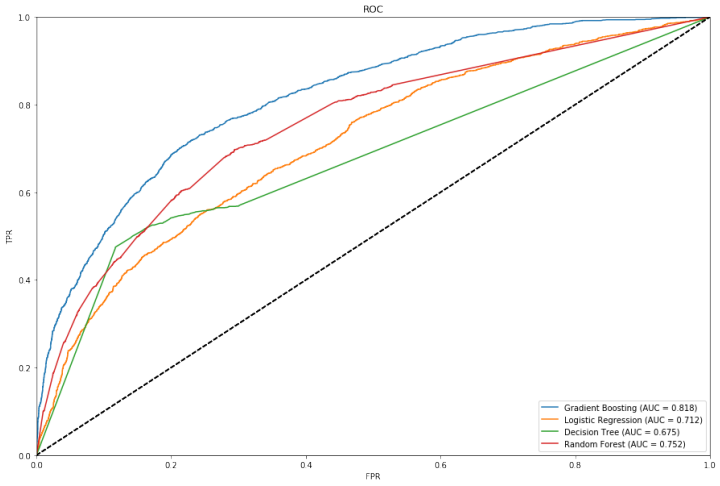

Fig. 6. ROC curve after feature reduction by feature importance

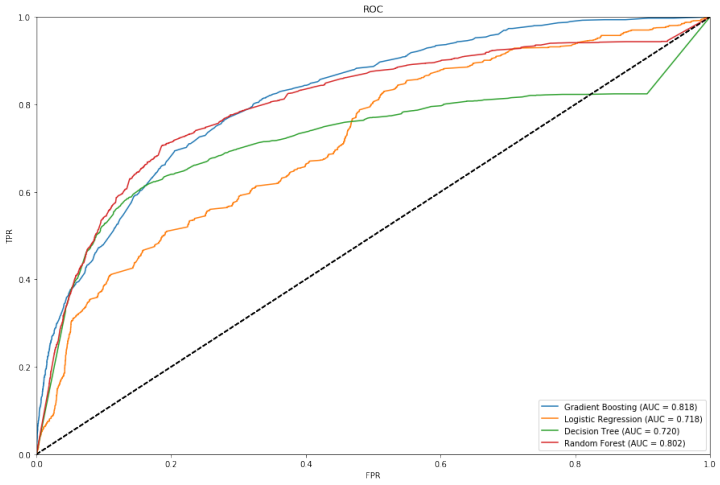

Fig. 7. ROC curve after feature reduction by mutual information

Based on the figures, it is observed that under the default setting, the decision tree and the random forest perform better after doing feature reduction by mutual information compared to the result coming from the full feature set. We think that this is because by default setting, 'min samples split' and 'min samples leaf' are set to 2 and 1 respectively. Thus, with a large feature set, the overfitting problem could be severe, but with a dataset only containing 4 features, overfitting is not too much of a problem. In comparing the two methods, it is clear that gradient boosting and logistic regression perform similar after doing both techniques, but decision tree and random forest perform better when doing

feature reduction by mutual information. These differences are explained by the fact that feature importance comes from the decision tree classifier, and is more conditional in its calculations. In addition, the optimal feature reduction method varies with respect to the classifier.

G. Results and Model Selection

1) Parameter/Feature Set Hypertuning: As specified previously, gridsearch was performed on each classifier with the full feature set, the feature set from feature importance, and the feature set from mutual information to determine which parameters should be used to obtain the best results (highest AUC score). For the gradient boost and the random forest we further used RandomGrid first, to narrow the range of parameter values to be passed to GridSearch, since GridSearch is very greedy on those algorithms.

The parameters which achieved the best results for decision tree were developed using the feature set reduced by mutual information. The results are shown below.

criterion	entropy
max_features	None
min_samples_leaf	10
min_samples_split	100

The parameters which achieved the best results for Logistic Regression were developed using the full feature set. The results are shown below.

C	solver	fit_intercept	class_weight
10	newton-cg	True	None

The parameters which achieved the best results for random forest were developed using the feature set reduced by mutual information. The results are shown below.

n_estimators	100
min_samples_split	12
min_samples_leaf	3
max_features	None
max_depth	20
criterion	entropy
bootstrap	True

The parameters which achieved the best results for gradient boosting were developed from the full feature set. The results are shown below.

loss	exponential
min_samples_leaf	30
min_samples_split	300
n_estimators	2000

In selecting the best model, we took a two-part approach. First, we used the parameters above to identify the model which maximized the AUC score on the validation set. Then, we developed a profit curve fitted to the training data and predicted on the validation set to determine which model maximized profits for the top 500 sales opportunities.

Figure 8 and Figure 9 show the ROC curves when using the best parameters for each classifier after parameter tuning on the full feature set and the feature set after feature reduction by mutual information. Logistic regression performs worse than the other classifiers in both graphs. Random forest and gradient boosting perform better than the other classifiers as their AUC scores are significantly higher. With higher thresholds, gradient boosting and random forest perform similar. However, gradient boosting out performs random forest at lower thresholds. Comparing the two graphs, it is clear that gradient boosting performs better with the full feature set, decision tree performs much better on the feature set after doing feature reduction and there's no significant difference between each setting for the other classifiers. Since gradient boosting with the full feature set has the best AUC score, we decide to stick with the full feature set and make profit curves and predictions based on that.

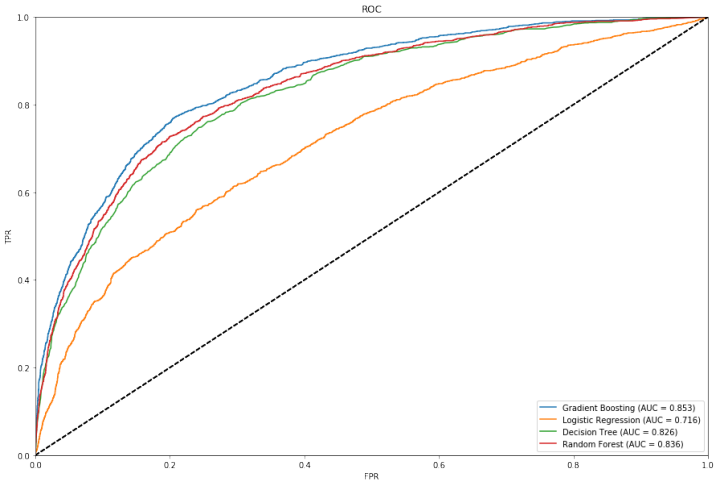

Fig. 8. ROC curve of models under best parameters on full feature set

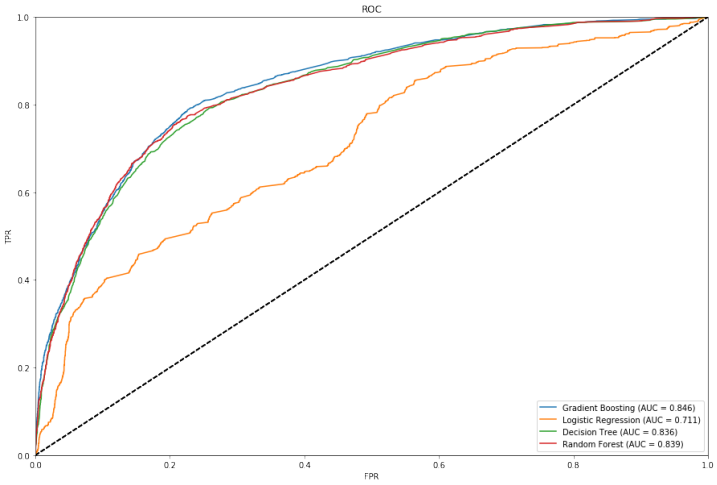

Fig. 9. ROC curve of models under best parameters on feature set after feature reduction by mutual information

Before developing the profit curves, we made the following assumptions about our business in order to estimate the cost of each opportunity:

- We are a company of which 40 employees are responsible for pursuing and fulfilling sales opportunities (38,400 employee hours over 6 months).
- It takes approximately 80 employee hours to pursue and fulfill each each sales opportunity (fixed cost analysis).
- So based on this, we can process approximately 480 sales over a 6-month period. We round up to 500 sales for simplicity.
- Thus, we are interested in targeting 500 sales opportunities over the course of six months.
- We assume that each worker earns $125/hour.

In developing a profit curve, we estimate the cost of each instance. Based on the assumptions above, we first assumed a fixed cost of $10,000 per instance. Figure 10 shows the profit curve when this fixed cost is assumed for all opportunities. Notice that we add a random classifier here which assigns a random possibility of success for each opportunity. Also, the black dashed line shows the cumulative cost for pursuing opportunities and the red dashed line shows our target 500 sales. We could see that in this case, Gradient Boosting will offer the most profit. And another interesting note here is that even if the company invests on opportunities randomly, it could still earn some money.

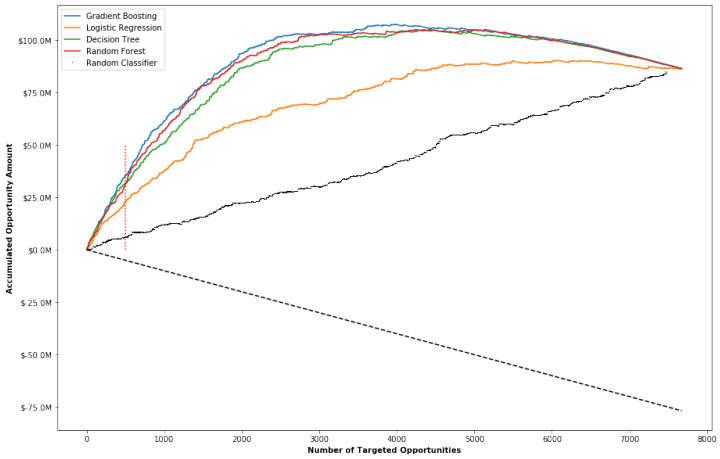

Fig. 10. Profit curve with cost = 10000

If we increase the cost, the company will earn less money until the random classifier will start to make the company lose money. Figure 11 shows a profit curve when we set the fixed cost for each sale as $30,000. We could see that in this case, the gradient boosting also offers the best profit.

At this point, we start to think that simply ranking the probabilities of success of each sale may not be the best ranking method. Because different sales have different opportunity amounts, we developed a weighted ranking (rank by winning probability * opportunity amount) to see

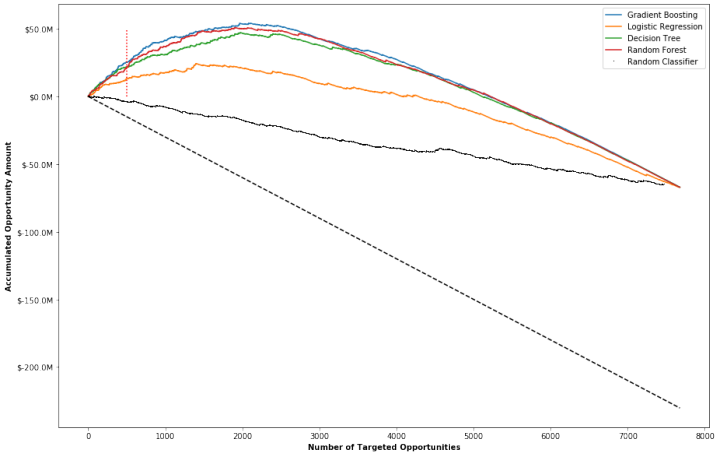

Fig. 11. Profit curve with cost = 30000

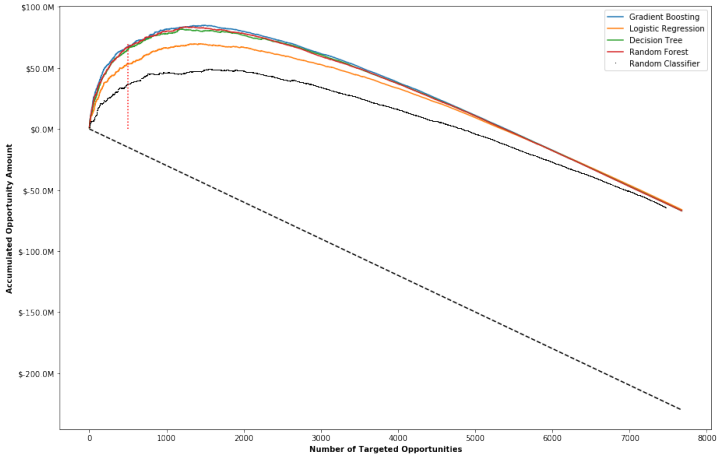

Fig. 13. Weighted Ranking Profit curve with cost = 30000

if applying this ranking could help the company earn more profit. Figure 12 and Figure 13 show the profit curve when assuming costs of $10,000 and $30,000, respectively, after applying the new ranking method. Notice that if we apply this new ranking method to our random classifier, it would be similar to rank opportunity amount with some noise. Based on the plots, random forest and gradient boosting has the highest profit in both cases and more importantly, by using this new ranking method, the company has the potential to earn more profit. This analysis makes sense because the larger the opportunity amount, the less the profit would be affected by an assumed fixed cost.

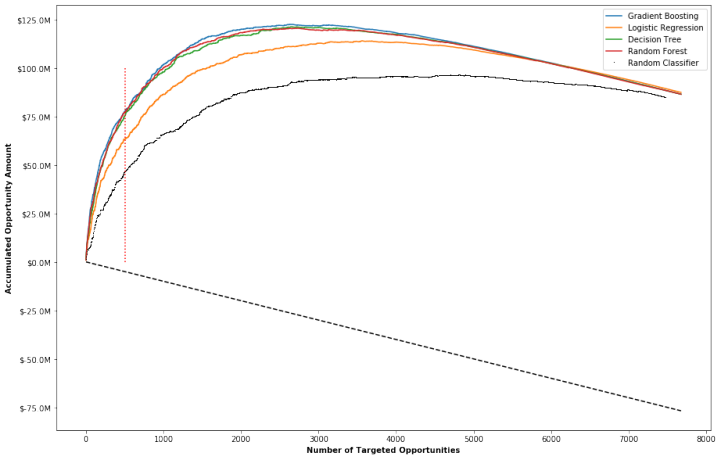

Fig. 12. Weighted Ranking Profit curve with cost = 10000

A realistic problem for the above profit curves is that we are assuming some fixed cost for all opportunities, but at the same time, some of the opportunity amount (i.e., possible revenue) could be less than the fixed cost. On the other hand, it is somewhat unrealistic for a company to invest the same amount of money for two sales with opportunity amounts of $5,000 and $50,0000. So a more realistic way to determine the cost for each sale is by defining a function like cost = some base rate + (a ratio) * opportunity amount. After applying this new method for cost, Figure 14 shows a profit curve with cost = $5,000 + 0.2 * opportunity amount. Notice that in this case, Gradient Boosting still shows the largest profit. If we apply the weighted ranking method as well as the new cost, the resulting profit curve is shown

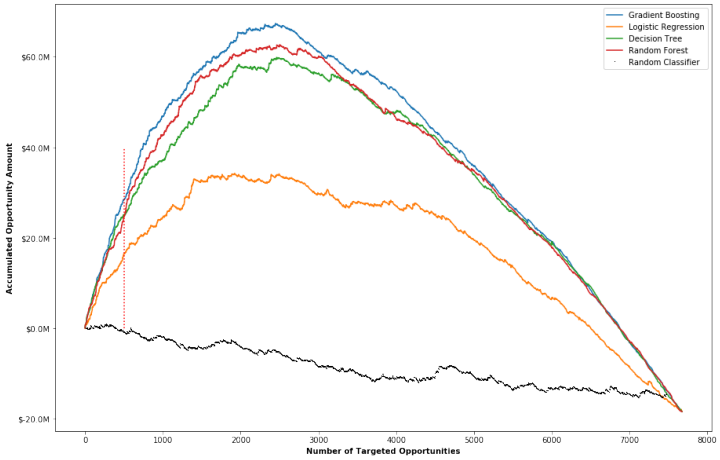

Fig. 14. Profit curve with cost = 5000 + 0.2 * opportunity amount

in Figure 15. Notice that Gradient Boosting still has the highest profit in both plots and the weighted ranking again increases the profit compared to Figure 14.

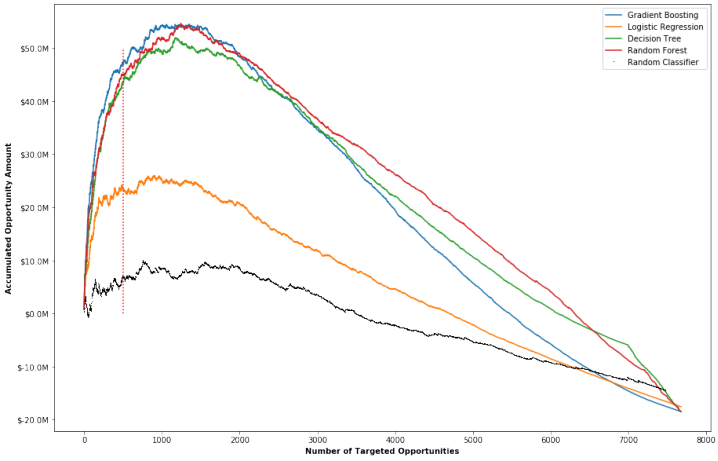

Fig. 15. Weighted Ranking Profit curve with cost = 5000 + 0.2 * opportunity amount

We notice that Gradient Boosting outperforms the other models in every aspect of the AUC analysis and the profit curve analysis, so we have selected this as our final model.

To test for generalization, first we defined the new training set as training set + validation set. Then, we fit our optimized gradient boosting classifier to the new training set. After this, we developed a ROC curve with AUC score and a profit curve based on predicting the values in the test set. The results are shown in Figures 16 and 17.

We determined that our model generalizes relatively well since the AUC is relatively equal from predicting the validation set to predicting the test set (0.853 vs. 0.854). This was surprising, because we expected the AUC to be a little lower in the generalization test. We double checked for data leakage, and the instances in the training, validation, and test sets are distinct and separate. We believe that because each set originated from the same initial data set of approximately 75,000 instances, this is why the model generalized so well. Therefore, it is recommended to

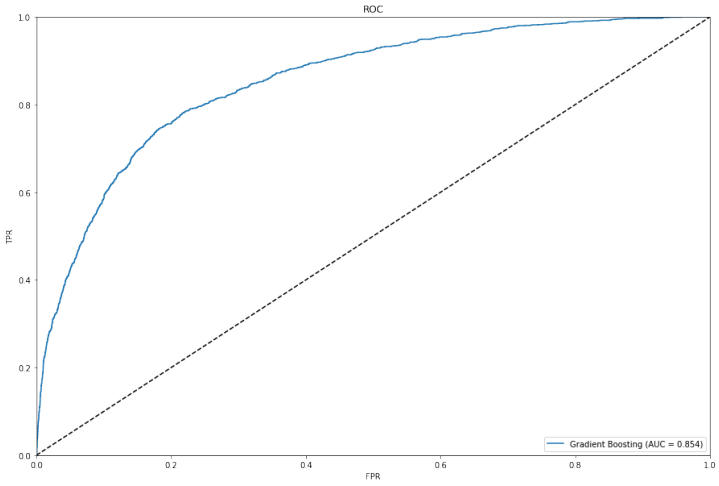

Fig. 16. Gradient Boosting AUC on test set

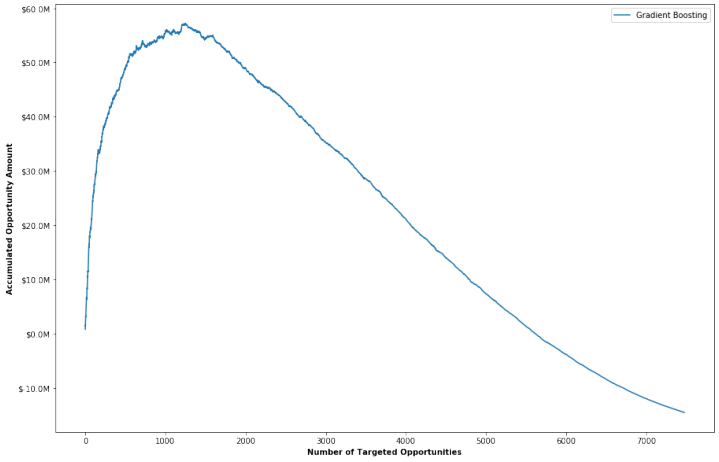

Fig. 17. Gradient Boosting profit curve on test set by using weighted ranking and (base cost = 5000, ratio = 0.2)

perform additional generalization tests with other datasets before the model is deployed.

VI. Deployment

We would use the fine-tuned gradient boosting model at our company to predict which sales opportunities to pursue. One idea is to make this model into a more user-friendly application that can be used by similar departments across the company to decide which sales opportunities to target. The user could input data such as ”number of employees”, ”cost per sale”, etc. as well as the historical data on which the model should be trained. We could have the software

retrain each time new historical data is inputted into the system by the user. As the data scientists of the company, each time the software is used is a good generalization test. Based on the results, we may re-calibrate the model and the resulting software as needed. Once we are consistently comfortable with the results, perhaps we could develop similar models and sell to companies in other industries who would like to more efficiently narrow down which sales opportunities to pursue. Some example sectors could include ferry retail suppliers, train retail suppliers and potentially others.

We identified no significant ethical considerations, as long as the historical data used to train the data is acquired legally. One concept we should consider is concept drift. If we notice that, for instance, the model does not perform well season to season, or during economic cycles, we could retrain the model each season, or for different stages in the economic cycle.

VII. Conclusion

In summary, we ran through several iterations of models to determine which worked best to determine the top 500 sales opportunities of an automotive supplier. To avoid overfitting, we ran through the following procedure to determine the best fit model. First, we removed features from the historical sales data which were not known at the time of the classification. Then, we split the data into training, validation, and test data sets based on 80%/10%/10% splits. We then iterated through classifiers, feature sets, and hyper parameters based on fitting to the training data set, and predicting the validation set. The gradient boosting classifier maximized the validation set AUC and profits (based on the profit curve analysis). Therefore, this was selected as our best model. To test for generalization, we then defined the new training set as the training set plus the validation set and fitted our gradient boosting classifier to the new training set. Then we calculated the AUC score and profit curve based on the test set using our best model as the final check for generalization. We proposed developing software for this model so that many people at our company can use it to target the top sales opportunities to pursue. Each time this occurs is a new opportunity for test for generalization, and possibly another opportunity to retrain the model, depending on the results. After some time, we plan to market this type of service to companies in other industries who look to select the best sales opportunities.

VIII. Appendix/Contributions

The code of this project can be found at: https://github.com/aspen8400/DS-GA-1001-proj.

Contributions:

- Alex: Wrote the sections I, II, VI and VII. Prepared the data and based on the data suggest to use random forest and gradient boosting. Suggested and made use of decision tree's feature importance to do feature reduction. Did parameter tuning for logistic regression and decision tree. Illustrated the business potential of the project.
- Jonas: Wrote the section III. Conducted the exploratory analysis (including plots) and implemented the random forest. Researched on the dataset and removed instances and features which should not be part of the project based on a deep understanding of the dataset (leakage). Suggested and made use of RandomGrid prior to Grid Search to save time. Did parameter tuning for random forest.
- Xiao: Wrote section V. Did data preprocessing and made base line models. Used Mutual Information as a way to do feature reduction. Did parameter tuning for gradient boosting. Made profit curves and other plots for different conditions.