Statistical Consulting Report

Addressing Class Imbalance in Random Forest Classifier used to Identify Potential Pancreatic Cancer Cases

Prepared by Lori Kolaczkowski

April 2021

Table of Contents

Abstract

Pancreatic ductal adenocarcinoma (PDAC) wouldn't be one of the deadliest cancers if an accurate screening test, simple enough to be as routine as measuring blood pressure, could be administered regularly to patients without any symptoms of PDAC, because once symptoms appear, it's already too late. This statistical analysis enhances the performance of a random forest classifier developed to execute such a screening test. The classifier reads measurements of components of urine which are strong indicators of pre-PDAC or early-stage PDAC, and classifies a patient as "low risk" or "elevated risk" for PDAC. A classification of "elevated risk" alerts doctors to the need for serious follow-up early enough to save the lives of those who may be silently developing the cancer. Similar screening tests exist but are less powerful and more invasive than a urine test, and thus are not given frequently enough. Class imbalance is identified as the cause of poor performance of the original classifier. To achieve performance goals required for approval and implementation of this test in real settings, a series of 10 models which correct for class imbalance are trained, tested, and compared. The prescribed model which meets the specified performance goals, is a random forest model using the SMOTE method to balance the response classes in the training set, with probability threshold tuned to 0.77.

Lori Kolaczkowski
Statistical Consulting Report, April 2021

Statement of Research

PDAC, the most common form of pancreatic cancer, and one of the deadliest and most aggressive of all cancers, has a very low survival rate with only 9% of patients with this cancer surviving 5 years beyond diagnosis. This poor survival rate is due to lack of detection in the early stages of development, when surgery can be effective in eliminating the cancer, as the cancer does not present symptoms until advanced. To detect this cancer early enough to substantially improve survivability, useful biomarkers associated with high risk of developing the disease must be identified through research and measured in routine screening tests which are clinically practical and ethical enough for widespread use. Traditionally, such biomarkers have been sourced from plasma or serum, requiring invasive collection and thus limiting frequency of screening.

This report provides requested expertise in statistical analysis needed for a relevant research study led by the client. The study investigates the viability of a completely noninvasive PDAC risk screening test, by examining the performance of patient age, sex, and a panel including creatinine and 3 urinary protein biomarkers, LYVE1, REG1B, and TFF1, in predicting elevated risk of PDAC. High levels of creatinine are associated with acute pancreatitis, a known risk factor for PDAC. LYVE1 (lymphatic vessel endothelial hyaluronan receptor 1) is a protein present in lymphatic vessels when pancreatic cancer is invading. REG1B (regenerating family member 1 beta) is a glycoprotein seen in patients with pancreatitis. TFF1 (trefoil factor 1) is a gastrointestinal secretory peptide which is more highly expressed during development of a variety of cancer types. LYVE1, REG1B, and TFF1 have each been found to be upregulated in PDAC precursor lesions, making them good biomarkers for pre-PDAC conditions and early-stage PDAC. If viable in its performance, a urine sample screening test based on these biomarkers would not only provide an inexpensive and noninvasive alternative method for early PDAC detection that is more conducive to frequent screening, but would likely be more effective given larger volume specimens containing higher concentrations of the biomarkers than what can be found in the blood.

The goal of the study is to train a prediction-focused classifier on this panel data collected from patients with and without PDAC, and classify test data with performance such that the classifier can primarily be expected to detect elevated risk for PDAC in a patient with such risk at least 90% of the time, and can be expected to label a patient without such risk as having risk no more than about 25% of the time (the latter being a secondary expectation).

Study Design

According to the client, the study is observational with limited knowledge of the data collection process. Patients with PDAC (n_1=44) and patients without PDAC (n_0=391), a total of 435 patients, were randomly selected to provide urine specimens. For each patient, biomarker presence was measured via enzyme-linked immunosorbent assays (ELISAs), creatinine was measured via ILab Aries analyzer, and these measurements were recorded along with age and sex of the patient. All variables were then creatinine-normalized. The urine specimens were collected and provided by various pancreatic cancer research centers and university laboratories across Europe, and they were proven to remain stable through the variable measurement process.
The data includes the binary categorical response variable, diagnosis ("1" if the patient who produced the urine sample had PDAC, "0" if the patient who produced the urine sample did not have PDAC), and the predictor variables sex ("F" if female and "M" if male), age (years), creatinine (mmol/L), LYVE1 (pg/ml), REG1B (pg/ml), and TFF1 (pg/ml). All predictors are continuous, except for the binary categorical variable, sex.

The classifier takes as input the predictor values for each sample, and outputs a class label for the patient corresponding to the levels of the categorical response variable, diagnosis – 0 = "no cancer", 1 = "PDAC". In a clinical setting, the class labels would instead refer to the patient's risk for PDAC - 0 = "low risk", 1 = "elevated risk". Patients who are labeled as having elevated risk would receive more invasive and expensive follow-up work such as imaging or pancreatic biopsy.

Using novice-level knowledge of machine learning algorithms, the client executed a random forest classification procedure on the data in R. After training and testing the classifier on this data, the client reported that the prediction performance was shown in the output to be 90% accurate, despite true PDAC cases being correctly classified as PDAC only 23% of the time.

Client Questions

01

Accuracy Metric

Why is the model output showing high accuracy, when it has such a high misclassification rate for cancer cases?

02

Low Performance

What is causing the classifier to perform poorly with respect to identifying cancer as cancer, and how do you recommend this problem be addressed in order to meet the goals of the study?

03

Importance of Predictor Variables

How can one get a sense of which variables are most important in predicting PDAC risk?

Statistical Method

The analysis conducted to answer the client's questions was carried out using the statistical software, R, and it serves to identify the cause of the client's undesirable results, develop a recommendation for changes to the client's analysis needed to achieve the study goals, and to demonstrate a way to determine variable importance. R code and output for the entire analysis are included in the appendix for the client to reference for variable importance and more. It began with obtaining the client's data and examining the dataset itself as well as visualizations created in performing exploratory data analysis (EDA). As the client suggested, the dataset was indeed complete and thoroughly cleaned. However, in confirming that the client defined the diagnosis level "1" (PDAC) to be the event of interest (sometimes termed the "success"), the levels of diagnosis had to be rearranged in R to ensure the program followed the same definition.

Exploratory Data Analysis

Considering the client's use of a random forest model, and a gradient boosting machine model considered later, EDA was not extensive. Both are nonparametric procedures carrying no assumptions for use in classification other than using a representative sample with a categorical response variable. Both are robust to outliers, and neither assume independent observations.
Since the client showed interest in variable importance, the analysis would include generation of variable importance plots, which can be misleading if any predictors are highly correlated with one another. Thus, a correlation heatmap (see Figure 1 below) was produced to check correlation levels between the continuous variables.

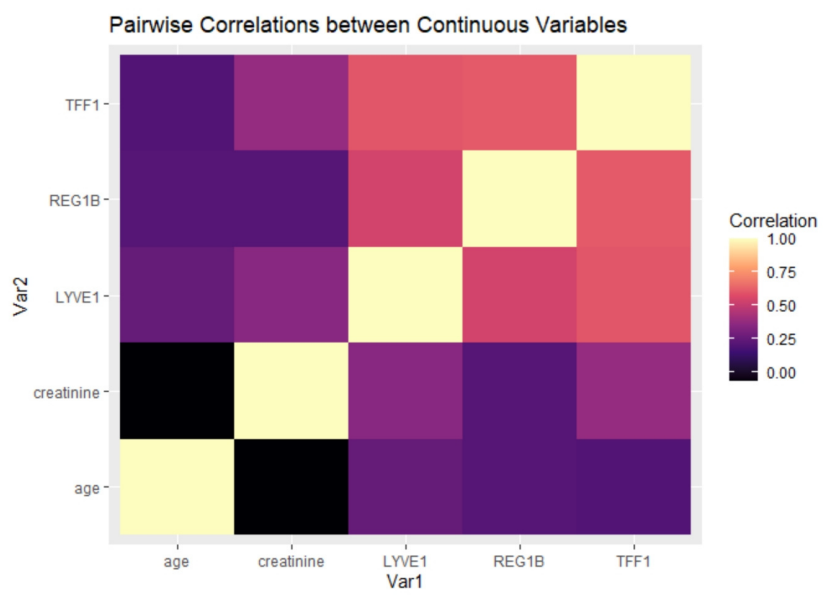


Figure 1. Heatmap showing pairwise correlations between continuous variables

Figure 1 shows no high correlation (> 0.70) was found between any unique pairs of variables. Instead, it shows low correlation between most of the continuous variables, and moderate correlation between all 3 pairs of biomarkers LYVE1, REG1B, and TFF1. And so, the variable importance plots to be generated later in the analysis were assumed to be trustworthy. Next, since the client's model is a classifier, it is important to see how well each class is represented in the data. To check the distribution of the response classes, a simple bar plot was created and is shown in Figure 2 below.

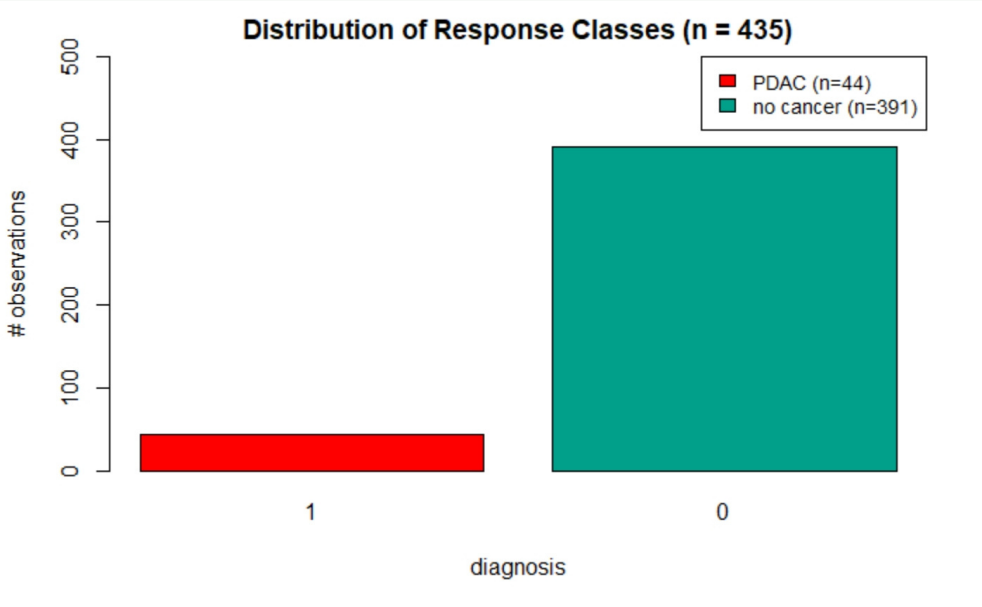


Figure 2. Bar plot showing imbalance among response classes

A moderate class imbalance is shown in Figure 2, for the response variable diagnosis. It is apparent that samples taken from patients without cancer outnumber samples taken from patients with PDAC, about 9:1. This certainly raised a red flag as class imbalance in the response variable often leads to poor predictions for the minority class if left unaddressed in the modeling process.

The Problem

In reviewing the client's methodology, it was found that this imbalance was indeed unaddressed. The response class imbalance, left unmitigated in the training of the classifier, provides reason for the classifier's poor performance in classifying the true cases of PDAC properly. Typical classification algorithms work to minimize error rate (maximize overall accuracy), without accounting for the class distribution. They are heavily biased towards the majority class. With only 44 samples representing PDAC, further reduced in the training set upon splitting the data into training and test sets, the algorithm tries to minimize error by classifying PDAC as non-cancer, resulting in a high misclassification rate within the PDAC class.

This leads to something called the Accuracy Paradox – the reason for the perceived discrepancy between accuracy value given in the output and misclassification rate in the minority class. The Accuracy Paradox occurs when a high proportion of the data is comprised of classes with negligible misclassification rates, while a small proportion of the data belongs to classes with high misclassification rates. Despite the classifier being terrible at predicting labels for the minority classes, overall, the classifier is great at making predictions, since most of the observations belong to the majority classes, and thus are classified correctly. This overall measure of classification performance is given in the testing output as "Accuracy", and is not the appropriate performance metric to use when the study is most concerned with classification performance of a minority class.

Before work can be done to mitigate the problems brought on by the class imbalance in the response, metrics are needed which will be most reliable in getting a good sense of how close the results are to the study goals. The study goals do not seek 90% overall accuracy, but rather a sensitivity of 0.90. These are completely different things, and thus accuracy should not be used as the performance metric because it is not indicative of distance to the goal. So, what is sensitivity and why should it be used as the performance measure for this study? Sensitivity is a confusion matrix metric which describes the true positive rate (in context of this study: proportion of times the classifier classifies a true cancer case to be cancer). Since the primary study goal is to detect PDAC when it is there at least 90% of the time, the goal is a sensitivity ≥ 0.9. But what else needs to be considered in measuring performance? The secondary goal of the study is to keep the proportion of non-cancer patients sent for unnecessary invasive and/or expensive follow-up tests below about 0.25. This corresponds to aiming for a false positive rate of about 0.25 or lower. The value for false positive rate is obtained by subtracting another confusion matrix metric, specificity, from 1.

Specificity is actually the true negative rate (proportion of times the classifier classifies a non-cancer case as a non-cancer case), and so subtracting it from 1 gives the false positive rate (proportion of times the classifier classifies a non-cancer case as a cancer case). Since this measures how close the results are to reaching the secondary goal of the study, it should be the secondary performance measure.

Ways to Mitigate

The following are some ways to handle imbalance among response classes in building a classifier.

- Get more data for the minority class until it is close in size to the majority class. This is by far the most ideal approach, as balance is achieved with genuine data, and as a result, the best predictive performance will be achieved. If it can be done, no need to consider any of the following mitigation steps.
- In randomly sampling for the training and test portions of the data, stratify the random sampling by response class to ensure the training and test data hold the same ratio of response classes. Without stratifying the split, most of the minority class observations could end up in the test set, leaving the classifier virtually untrained on the minority class, and resulting in very poor predictions for that class.
- Try integrating different balance-recovery sampling methods (described on next page) for the training data into the model fitting and validation process.
- Repeat third bullet point for each of a few different classification algorithms appropriate for the goals of the study.

In the case of this study, the client has stated that collection of more data is unfortunately not possible. Therefore, performance would be enhanced by using stratified random sampling in forming the training and test datasets, and then applying a balancing method to the training data. There are 3 ways to balance the response classes when you cannot collect more genuine data – majority reduction (under-sampling), minority augmentation (oversampling), or a mix of both.

Balancing Method	Description
Under-sampling (Figure 3, upper left)	Randomly sample (without replacement) observations from the majority class to use until it is even with the minority class in size.
Oversampling (Figure 3, top right)	Randomly sample observations (with replacement) of the minority class until it is even with the majority class, i.e., generate copies of original minority observations.
SMOTE *Synthetic Minority Oversampling TEchnique* (Figure 3, lower right)	Generate synthetic minority samples to augment the existing ones until even with the majority class, through randomly selecting minority class observations and, for each, interpolating new values between the selected observation and its minority neighbors.
ROSE *Random Over-Sampling Examples* (Figure 3, lower left)	Simulates a new balanced training set by producing synthetic samples for each class from their respective estimated distributions (yellow and purple ovals in Figure 3) based on the original data.

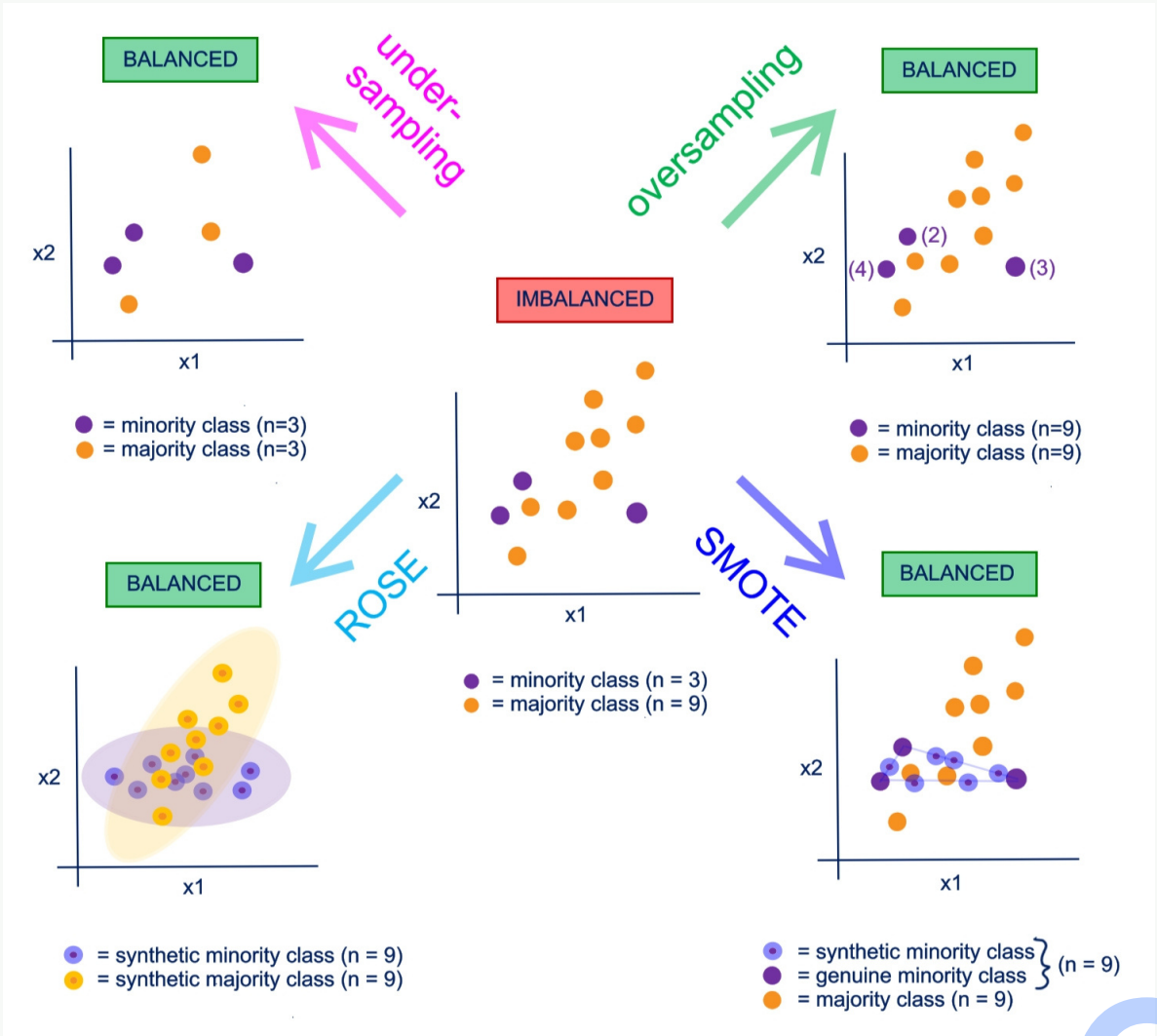


Figure 3. Visual comparison of methods for balancing classes of the training set

Analysis

The approach taken with the analysis was to try several models, using different combinations of algorithm type and balancing method type, since there is no one-size-fits-all solution. The approach with the best results changes from one dataset to the next, and with changing study goals. The classification performance of each were compared and the model(s) with highest sensitivity further tuned to optimize performance with respect to goals for sensitivity and 1-specificity.

The algorithms compared were limited to random forest (RF) and gradient boosting machines (GBM), since these generally yield the highest predictive performance and loss of interpretability is not an issue with prediction being the sole concern of the study. It made sense to include GBM since it is particularly well-suited for high bias/low variance procedures, which is an expected result of reducing training data in using the under-sampling balancing technique. And RF is particularly well-suited for low bias/high variance procedures expected to result from augmenting training data via oversampling techniques. There are more balance-recovery sampling techniques than under-sampling, oversampling (via minority resampling), SMOTE, and ROSE, but these were selected as balancing tools in this analysis due to their popularity in the machine learning community, and availability of packages in R for deployment. To adhere to the client's definition of PDAC as the event ("success"), this definition was maintained for every model. The primary performance metric chosen was sensitivity, with a goal of sensitivity greater or equal to 0.90, and the secondary metric chosen was 1-specificity, with a goal of 1-specificity no greater than about 0.25. These metrics and limits were chosen to conform to the goals of the study as previously discussed. For this analysis, no variables were transformed since normalization and transformations are not needed for the nonparametric models used. No variables were removed since these models are unaffected by multicollinearity if it exists, and each variable helps to predict PDAC risk to some extent, helping to maximize predictive performance which is the only concern in this study.

To begin, the client's RF model was reproduced for comparison with balanced models. The client's model used a training set and test set assembled with random sampling (not stratified by class), nor did it make use of a balancing technique. Next, 4 balanced RF models were produced, each using a unique balancing method. All balanced models used a training set and test set assembled via stratified (by class) random sampling of observations from the original dataset, for the reason given in "Ways to Mitigate", and all balanced models were trained and tested using this same training set and test set. Regardless of use of stratified random sampling, training sets for all models used 70% of the original data. Each model was fit on its training set using repeated K-fold cross-validation, with K=10 folds and 10 repeats. For balanced models, the balance sampling method was deployed on the imbalanced training set within each cross-validation fold, independently. This served to ensure model performance values were robust, with or without balancing. Some may wonder why training and test sets were used for these algorithms rather than using out-of-bag (OOB) samples, and the reason is that use of training and test sets allow for integration of the balancing methods in cross-validation.

For each model, its test data was then fed to the model, and probabilities associated with each class for all observations were output by the model. The probabilities for each observation were subjected to the following rule and converted to a class label (or class prediction), where tau (τ) = 0.5 is the default tuning parameter value (the probability threshold) of the classifier used in all models, initially, and later tuned to different values for selected models:

> If the probability that the observation = "no cancer" is $\geq \tau$, classify the observation as "no cancer". Otherwise, classify the observation as "PDAC".

This was repeated using the GBM algorithm in place of RF, generating 4 additional balanced models. For comparison's sake a GBM model using the same unstratified training and test sets used in the client's model was produced without use of any balancing technique. The 10 models examined in this analysis are listed below.

RF/imbalanced (client model)	GBM/imbalanced
RF/balanced (under-sampling)	GBM/balanced (under-sampling)
RF/balanced (oversampling)	GBM/balanced (oversampling)
RF/balanced (SMOTE)	GBM/balanced (SMOTE)
RF/balanced (ROSE)	GBM/balanced (ROSE)

For each model, the predicted classes were then compared to the actual classes and a confusion matrix was produced. Confusion matrix metrics, sensitivity and specificity, for each model of a given algorithm type, were visualized in a single plot to avoid confusion and render the results more easily digestible. These metrics served as model validation as they show whether or not the model is performing well enough to meet study goals. The model with highest sensitivity within the RF group, and the model with the highest sensitivity within the GBM group were then compared with respect to sensitivity and 1-specificity upon tuning τ to a series of different values. Tuning τ further optimizes the results towards goal values. This was visualized with ROC plots in order to see which of the 2 tuned models were best for achieving the study goals. To accommodate the client's interest in variable (feature) importance, feature importance plots were made for the two competing models.

Results

As shown in the left panel of Figure 4 below, the under-sampling method (under) gave the highest sensitivity (0.77) out of all 5 GBM models. The right panel of the same figure shows the SMOTE method resulted in the highest sensitivity among the 5 RF models, however, it tied with the GBM/under-sampling model at a sensitivity of 0.77. So the best RF model and the best GBM model, before tuning τ, can only be expected to detect PDAC 77% of the time, 13 percentage points below the study goal of 90%. The analysis goes forward considering only these 2 models, and tunes τ in order to reach the study goals It is interesting to note that all balanced models outperformed the imbalanced models for both RF and GBM, with an increase of almost 0.4 in sensitivity between imbalanced GBM and GBM/under-sampling, and an increase of almost 0.55 in sensitivity between imbalanced Client RF and RF/SMOTE.

Sensitivity ≥ 0.90	GOALS	1-Specificity ≤ 0.25

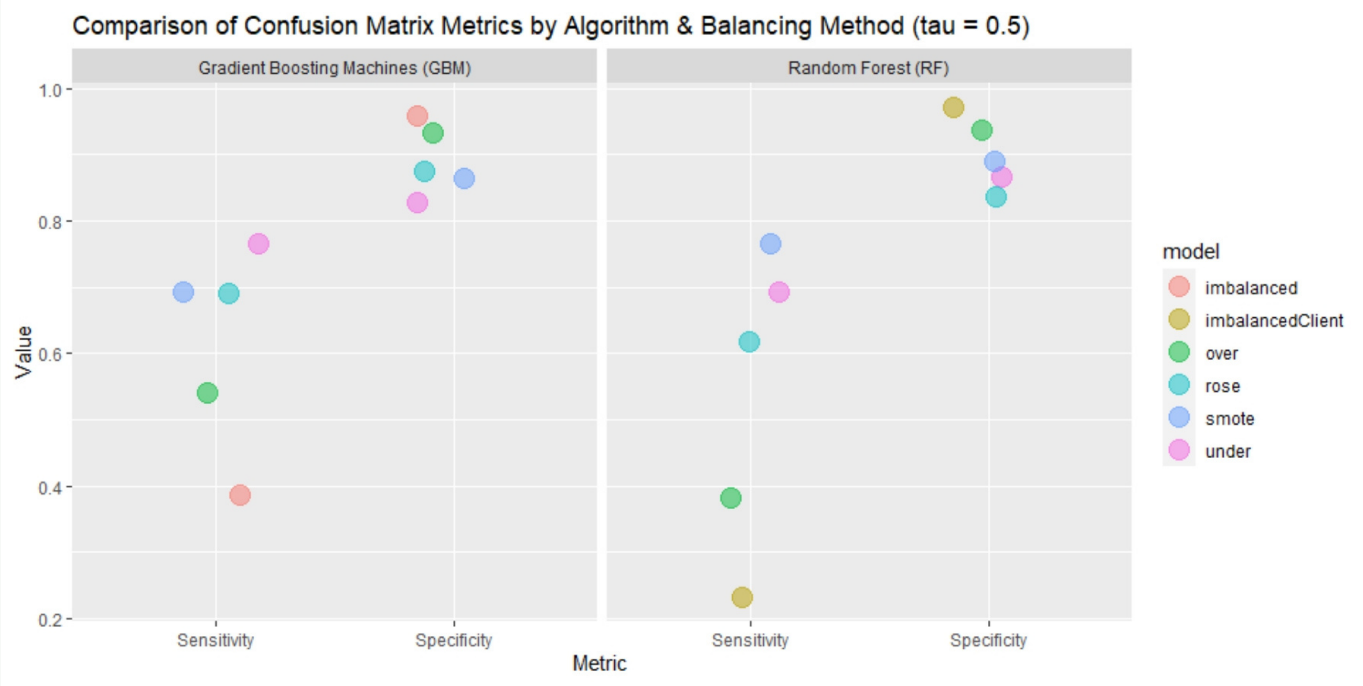


Figure 4. Pre-tuning comparison of the 10 models, by algorithm type, on goal metrics. Note that sensitivity is the primary focus, and though some models are much closer to the goal of 0.90 than others, none of them make it there just yet.

Relative feature importance for the GBM/under-sampling model is given on the left in Figure 5 below, with feature importance for the RF/SMOTE model on the right. In either model, it is clear LYVE1 and age are the strongest predictors for PDAC risk, with REG1B in 4th place, and sex being the weakest predictor (though it still adds value for prediction). The rank of TFF1 and creatinine as predictors depends on the model.

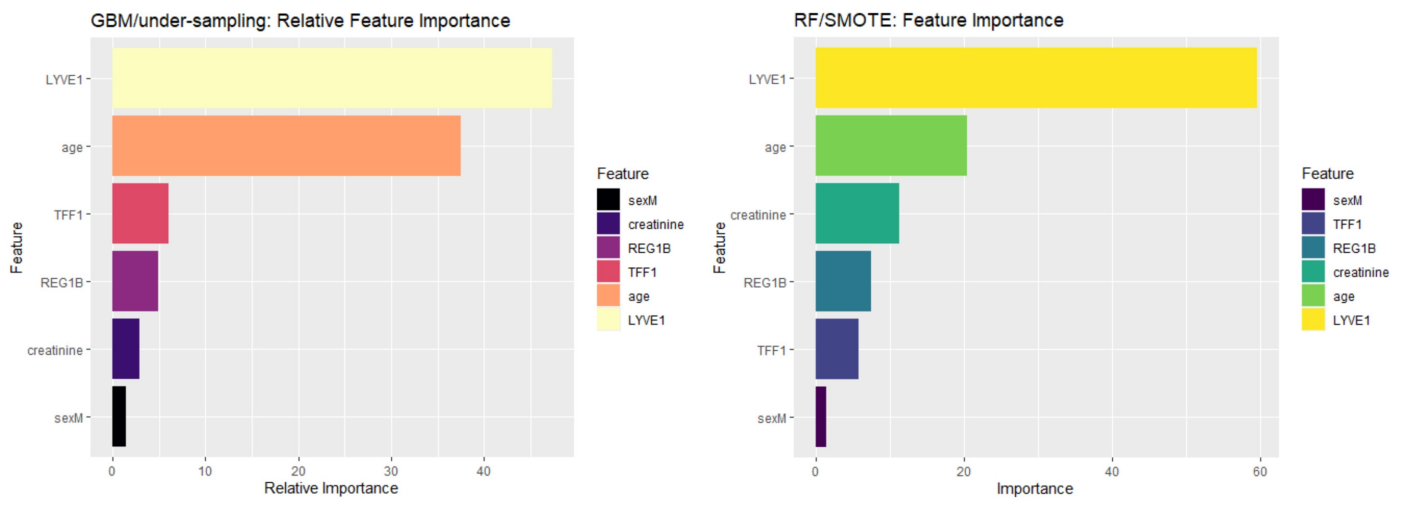


Figure 5. Feature Importance for RF/SMOTE and GBM/under-sampling

Returning to the model results on the previous page, sensitivity must be increased through tuning threshold values. The ROC curve for the RF/SMOTE model is shown below in Figure 6, and it shows the sensitivity and specificity achieved at different threshold values (the values immediately left of the "Spec" and "Sens" values). The threshold, τ, simply dictates how much evidence (probability) is required that a patient does not have cancer in order to classify a patient as "no cancer". In order to increase the percentage of time the classifier detects PDAC when a patient has it, τ has to be increased to require stronger evidence that a patient does not have cancer to be classified as such. Tuning τ upward from 0.5 to 0.77 or higher allows for the sensitivity goal of ≥ 0.90 to be achieved. However, for the 1-specificity goal of $\leq$ about 0.25 (or equivalently specificity $\geq$ about 0.75) to be essentially met simultaneously, τ cannot exceed 0.77. Therefore, a τ of 0.77 is optimal for the RF/SMOTE model, carrying sensitivity of 0.92 and 1-specificity of 0.26 (later a table of more precise thresholds implies this is more likely to be 0.25). The ROC curve for the GBM/under-sampling model is shown below in Figure 7, and it shows the sensitivity and specificity achieved at different threshold values. Tuning the probability threshold (τ) upward from 0.5 to 0.66 or higher allows for the sensitivity goal of ≥ 0.90 to be achieved. However, for the 1-specificity goal of $\leq$ about 0.25 (or equivalently specificity $\geq$ about 0.75) to be essentially met simultaneously, τ cannot exceed 0.66. Therefore, a τ of 0.66 is optimal for the GBM/under-sampling model, carrying sensitivity of 0.92 and 1-specificity of 0.26. This appears to exactly match the performance of the RF/SMOTE model at $\tau = 0.77$, though this is due to rounding. RF/SMOTE actually has a very slightly lower value for 1 - specificity. The AUC (area under the curve) is a type of overall measure of performance and is included as secondary support for the final model choice.

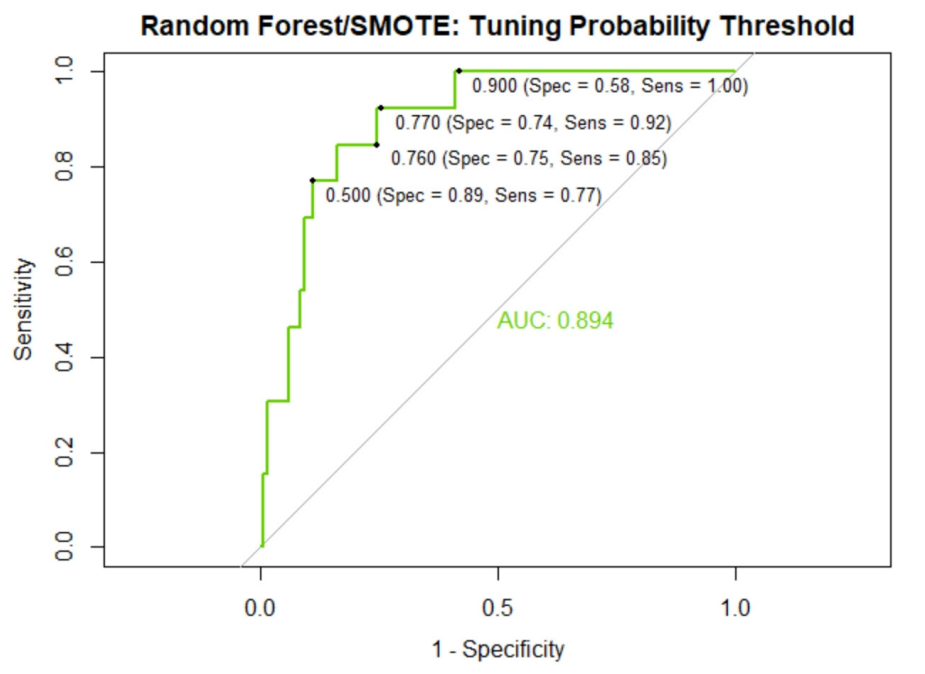


Figure 6. ROC plot for RF/SMOTE model

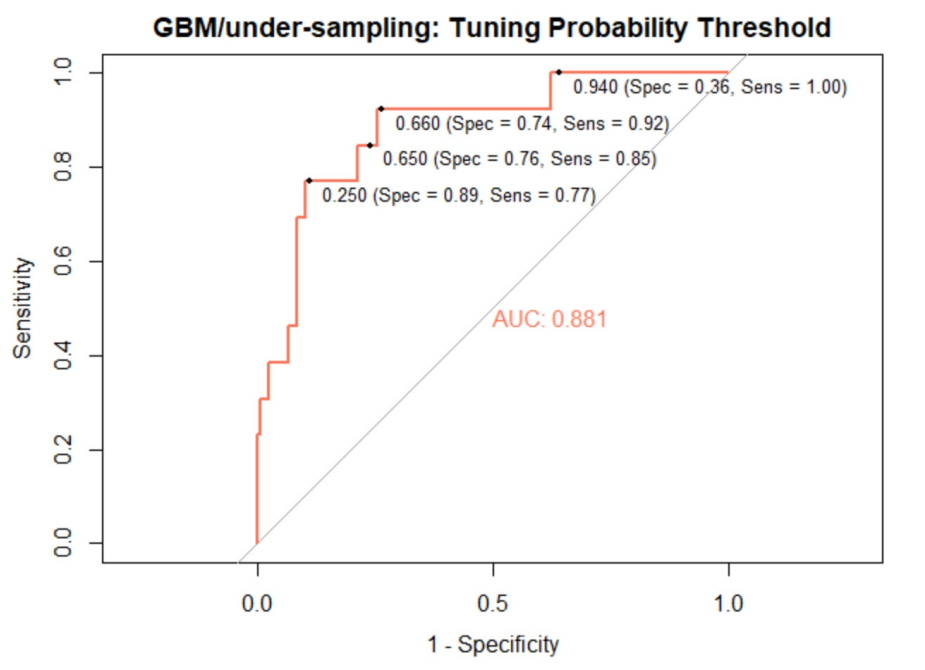


Figure 7. ROC plot for GBM/under-sampling model

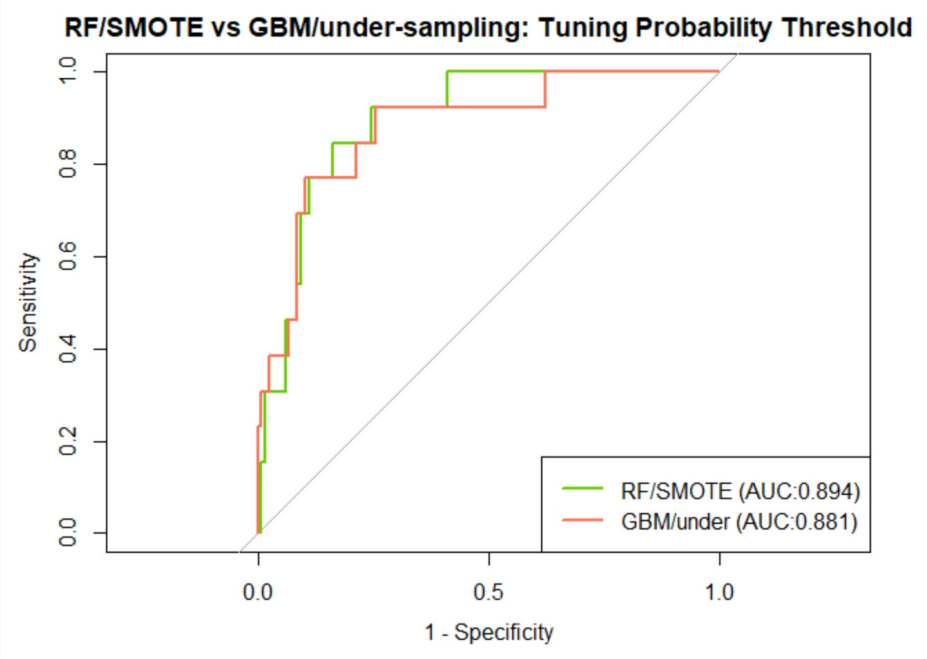


Figure 8. ROC plots for RF/SMOTE and GBM/under-sampling together

Figure 8 above helps to compare the ROC curves of the competing models. The sensitivity/specificity results are very close for the two models most of the time, but 1- specificity can be substantially different in certain ranges of sensitivities. For instance, for sensitivities between 0.92 and 1.00, RF/SMOTE achieves significantly lower false positive rate (1-specificity) than GBM/under-sampling – 0.42 for RF/SMOTE, and 0.64 for GBM/under-sampling. That is, if study goals were to change, and sensitivity needed to be > 0.92, RF/SMOTE would be much preferred to GBM/under-sampling as, given the same true positive (PDAC detection) rate of 0.92, the false positive rate for RF/SMOTE would be 0.22 lower than that of GBM/under-sampling. And what is preferred for this study is for false positive rate to be minimized as much as possible while the sensitivity remains above 0.90. Considering the current study goals, RF/SMOTE with $\tau = 0.77$ is really quite close (though slightly more optimal in 1-specificity value as shown above in Figure 8) to GBM/under-sampling with $\tau = 0.66$, and both satisfy the stated goals of the study. RF/SMOTE also has a slightly higher AUC (better overall performance).

Since RF/SMOTE appears to be slightly more optimal, a more granular breakdown of sensitivity and 1-specificity at various threshold values for the model are given below in Figure 9. Midpoint interpolation between the values for thresholds 0.767 and 0.782 suggests a value of 0.254 for 1-specificity at a threshold of 0.77. Thus, a threshold of 0.77 for the RF/SMOTE model does indeed achieve the study goals of sensitivity ≥ 0.90, and 1-specificity ≤ about 0.25.

RF/SMOTE: Comparing Probability Thresholds

Threshold (Tau)	Sensitivity (Prop. of true +)	1 - Specificity (Prop. of false +)
0.913	1.000	0.436
0.909	1.000	0.427
0.903	1.000	0.419
0.898	1.000	0.410
0.895	0.923	0.402
0.892	0.923	0.393
0.887	0.923	0.385
0.881	0.923	0.368
0.877	0.923	0.359
0.875	0.923	0.350
0.873	0.923	0.342
0.871	0.923	0.316
0.869	0.923	0.299
0.859	0.923	0.291
0.845	0.923	0.282
0.832	0.923	0.274
0.810	0.923	0.265
0.782	0.923	0.256
0.767	0.923	0.248
0.763	0.846	0.248
0.750	0.846	0.231

Figure 9. Table of thresholds and corresponding metrics of interest for RF/SMOTE

Figure 10 below visually compares the sensitivity and specificity of the RF/SMOTE model before tuning (smote), and the same model after tuning τ (tau) to 0.77 (smoteTuned).

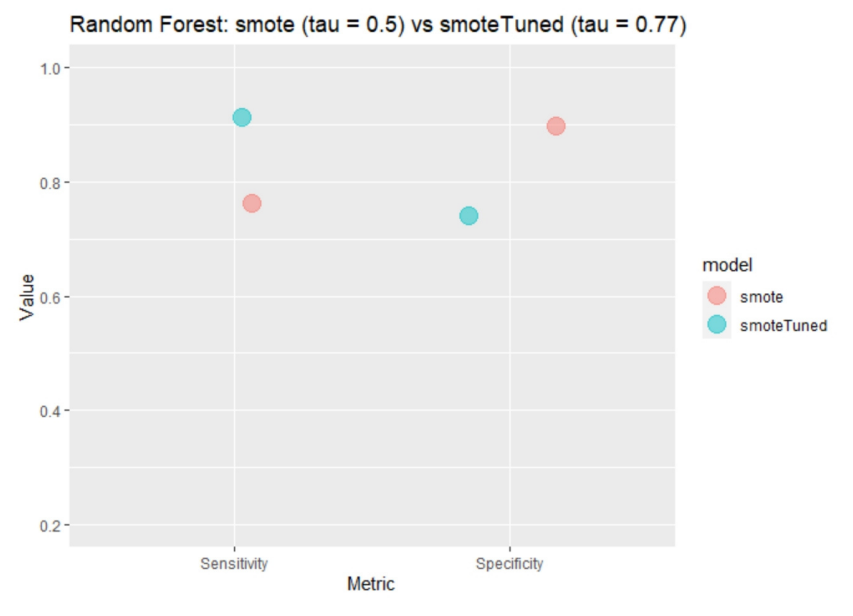


Figure 10. Plot of RF/SMOTE before and after tuning

The RF/SMOTE classification model with τ = 0.77 can be expected to detect cancer risk in patients more than 92% of the time when cancer risk is there, giving good reason to continue with more invasive and/or expensive procedures such as biopsy of the pancreas and imaging, which will catch presence of PDAC early enough to save many lives. And it comes with a relatively low cost in terms of false positives – only 25% of those without high cancer risk will be recommended to undergo further follow-up procedures.

Conclusions

Answers to the client's questions:

01 Accuracy Metric

The reported performance measure discrepancy was due to using a performance metric which is not appropriate for the goals of the study. Given the goals of this study, total accuracy should not be the chosen metric, but rather sensitivity should be the primary metric with 1-specificity as the secondary metric.

02 Low Performance

The initial poor performance of the random forest classifier in detecting cancer was a result of moderate class imbalance in the response variable. In integrating methods for balancing the response classes into the model training process as detailed in the methods section, the following classification procedure not only meets the specified goals, but will also be more robust to any changes in study goals, and is thus recommended for use in the study:

> Random Forest algorithm using SMOTE with classification cutoff $\tau = 0.77$.

Although this procedure meets the study goals, it could have shown improved results if the study had allowed for collection of enough genuine urine samples from subjects with PDAC to balance with the number of urine samples from subjects without PDAC. Still, the above approach is an effective alternative.

03 Importance of Predictor Variables (Features)

Feature importance can be produced in R from random forest models using varImp(model) from the caret package and from gradient boosting machine models using summary(model), and both can be plotted using ggplot2 as shown in the appendix under "Check variable importance...".
But be aware that high correlation between features can result in misleading feature importance plots. Correlations can be checked by making a correlation heat map as shown in the appendix in the exploratory data analysis portion.

Closing Statements

It can be done – even with moderate class imbalance, pancreatic cancer risk can be detected accurately with a simple noninvasive and inexpensive urine sample test by running the resulting data through the recommended classification procedure. And its predictive performance, since it meets goals for true positive rate and false positive rate, renders the procedure clinically practical enough to make implementation of the urine sample screening test a feasible and attractive option in real settings.

Many people die from PDAC because it must be caught very early to beat it, but it instead advances undetected for lack of strong enough evidence to warrant the invasive and/or costly tests which are needed to find the cancer. Run through an appropriately crafted classifier, the biomarkers measured in this study's urine sample panel are strong enough predictors of PDAC to provide the means to routinely and accurately screen patients for follow-up need, long before PDAC can become advanced enough to present the first signs and symptoms that it's already too late.

As other urinary proteins and peptides have shown promise as biomarkers for other cancers including lung cancer, colon cancer, and ovarian cancer, this study can inform similar studies in other cancer areas in which such proteins are explored as panel candidates for simple and accurate routine screening tests. This study should also serve as an example of some ways to handle class imbalance in classification procedures, and to encourage sampling designs which allow for balanced classes in categorical responses when classifiers are to be trained on them, regardless of the research area.

Appendix

R Code and Output for Client's Reference

Analysis Script & Output: Addressing Class Imbalance in Random Forest Classifier used to Identify Potential Pancreatic Cancer Cases

Lori Kolaczkowski (kolaczkowskil@stat.tamu.edu)

4/9/2021

```
#load packages
library(ggplot2) #for some plots
library(viridis) #for plot color
library(wesanderson) #for plot color
library(reshape2) #for heat map (melting correlation)
library(caret) #for all the ML algorithms
library(DMwR) #for smote
library(ROSE) #for rose
library(dplyr) #for data manipulation needed for confusion matrix plots
library(tidyr) #for data manipulation needed for confusion matrix plots
library(pROC) #for ROC plots
library(knitr) #for ROC table
```

```
#read in the data
data <- read.csv("D:/STAT 684 - Consulting/Project/pancCancer.csv", header = TRUE)

#check it out (un-comment next line to print the data)
#data
str(data)
```

```
## 'data.frame':    435 obs. of  7 variables:
##  $ age       : int  33 81 51 61 62 53 70 58 59 56 ...
##  $ sex       : chr  "F" "F" "M" "M" ...
##  $ diagnosis : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ creatinine: num  1.832 0.973 0.78 0.701 0.215 ...
##  $ LYVE1     : num  0.89322 2.03758 0.14559 0.0028 0.00086 ...
##  $ REG1B     : num  52.9 94.5 102.4 60.6 65.5 ...
##  $ TFF1      : num  654.3 209.5 461.1 142.9 41.1 ...
```

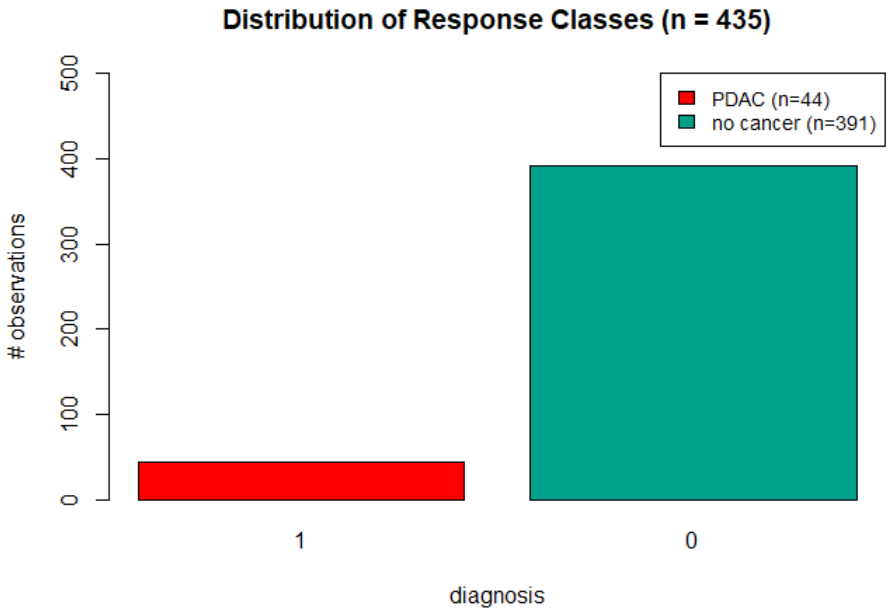
```
#convert categorical variables to factors
data$sex <- as.factor(data$sex)
data$diagnosis <- as.factor(data$diagnosis)

#reorder levels of diagnosis so that "1" shows as 1 and "0" shows as 2
#by default caret will call the 1st level the event (or success)
data$diagnosis <- factor(data$diagnosis, levels = c("1","0"))
str(data)
```

```
## 'data.frame':    435 obs. of  7 variables:
##  $ age       : int  33 81 51 61 62 53 70 58 59 56 ...
##  $ sex       : Factor w/ 2 levels "F","M": 1 1 2 2 2 2 2 1 1 1 ...
```

```
##  $ diagnosis : Factor w/ 2 levels "1","0": 2 2 2 2 2 2 2 2 2 2 ...
##  $ creatinine: num  1.832 0.973 0.78 0.701 0.215 ...
##  $ LYVE1     : num  0.89322 2.03758 0.14559 0.0028 0.00086 ...
##  $ REG1B     : num  52.9 94.5 102.4 60.6 65.5 ...
##  $ TFF1      : num  654.3 209.5 461.1 142.9 41.1 ...
```

First, always run through some exploratory data analysis.

```
###### check distribution of predictors/features

# check distribution of response/supervisor
barplot(table(data$diagnosis),
        ylab = "# observations",
        xlab = "diagnosis",
        ylim = c(0,500),
        main = "Distribution of Response Classes (n = 435)",
        col = wes_palette("Darjeeling1", 5, type = "continuous"))
legend('topright', legend=c("PDAC (n=44)", "no cancer (n=391)"), cex=0.9,
       fill = wes_palette("Darjeeling1", 5, type = "continuous"))
```

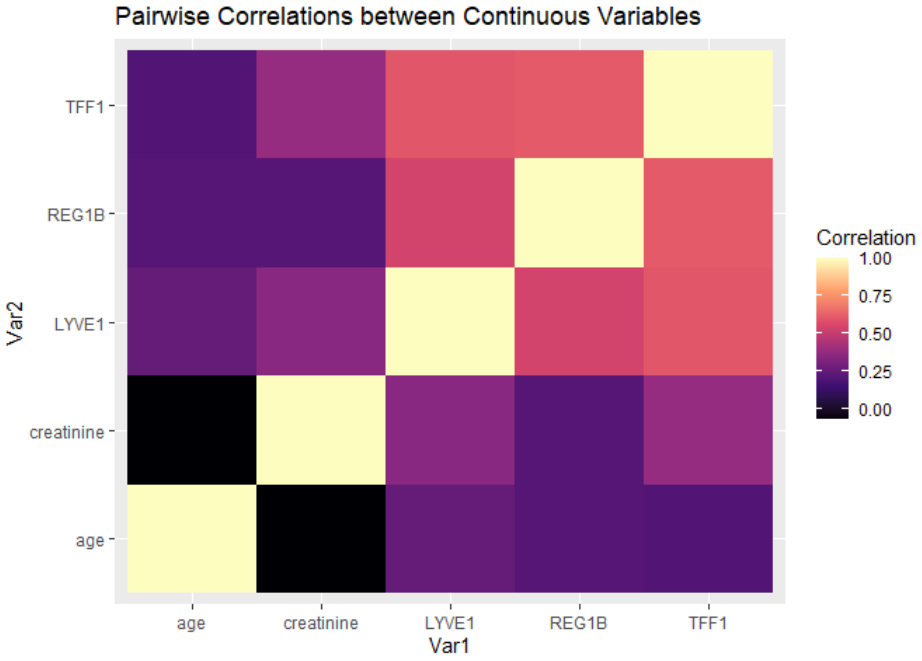



```
###### check correlation between all pairs of features with a correlation heatmap
# make and melt the correlation matrix for numeric variables
feats <- data[,-c(2,3)]
cormat <- round(cor(feats),2)
melted_cormat <- melt(cormat)
melted_cormat
```

```
##          Var1       Var2 value
## 1         age        age  1.00
## 2  creatinine        age -0.07
## 3       LYVE1        age  0.25
## 4       REG1B        age  0.21
## 5        TFF1        age  0.20
## 6         age creatinine -0.07
## 7  creatinine creatinine  1.00
```

```
## 8       LYVE1 creatinine  0.35
## 9       REG1B creatinine  0.21
## 10       TFF1 creatinine  0.38
## 11        age      LYVE1  0.25
## 12 creatinine      LYVE1  0.35
## 13      LYVE1      LYVE1  1.00
## 14      REG1B      LYVE1  0.54
## 15       TFF1      LYVE1  0.60
## 16        age      REG1B  0.21
## 17 creatinine      REG1B  0.21
## 18      LYVE1      REG1B  0.54
## 19      REG1B      REG1B  1.00
## 20       TFF1      REG1B  0.61
## 21        age       TFF1  0.20
## 22 creatinine       TFF1  0.38
## 23      LYVE1       TFF1  0.60
## 24      REG1B       TFF1  0.61
## 25       TFF1       TFF1  1.00
```

```
# make the correlation heatmap using viridis palettes
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  scale_fill_viridis_c(option="magma", name="Correlation") +
  labs(title="Pairwise Correlations between Continuous Variables")
```



Original random forest model:

No stratified train/test split, and no sampling to correct imbalance

```
# data is split into train and test sets using random sampling (not stratified by response class),
#so that 70% of the data will be used to train the model
set.seed(777)
these_client <- sample(1:nrow(data), floor(nrow(data)*0.7))
train_client <- data[these_client, ]
test_client <- data[-these_client, ]
```

```
# repeated K-fold cross-validation object is created (K=10, repeats=10)
cv_client <- trainControl(method = "repeatedcv",
                  number = 10,
                  repeats = 10,
                  verboseIter = FALSE)

# RF model is fit implementing above CV object
set.seed(777)
modelrf_client <- caret::train(diagnosis ~ .,
                           data = train_client,
                           method = "rf",
                           trControl = cv_client)

final_client <- data.frame(actual = test_client$diagnosis,
                       predict(modelrf_client, newdata = test_client, type = "prob"))
final_client$predict <- ifelse(final_client$X0 > 0.5, 0, 1)
final_client$predict <- as.factor(final_client$predict) #predictions must also be a factor
cmRF_imbalancedClient <- confusionMatrix(final_client$predict, test_client$diagnosis)
```

```
## Warning in confusionMatrix.default(final_client$predict, test_client$diagnosis):
## Levels are not in the same order for reference and data. Refactoring data to
## match.
```

```
cmRF_imbalancedClient
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   0
##          1   3   3
##          0  10 115
##
##                Accuracy : 0.9008
##                  95% CI : (0.8363, 0.9461)
##     No Information Rate : 0.9008
##     P-Value [Acc > NIR] : 0.57314
##
##                   Kappa : 0.27
##
##  Mcnemar's Test P-Value : 0.09609
##
##             Sensitivity : 0.23077
##             Specificity : 0.97458
##          Pos Pred Value : 0.50000
##          Neg Pred Value : 0.92000
##              Prevalence : 0.09924
##          Detection Rate : 0.02290
##    Detection Prevalence : 0.04580
##       Balanced Accuracy : 0.60267
##
##        'Positive' Class : 1
##
```

```
## Printing cmRF_original shows true cancer cases misclassified 10/13 times (90% total accuracy)
```

Mitigation Suggestions

Let's be sure to stratify the train/test split

```
# createDataPartition carries out random sampling of observations, stratified by response class.
#70% of the data will be used to train the model
set.seed(777)
these <- createDataPartition(data$diagnosis, p = 0.7, list = FALSE)
train <- data[these, ]
test  <- data[-these, ]
```

Now let's fit a series of random forest models - each implementing a different sampling method (undersampling, oversampling, SMOTE, ROSE) to achieve class balance in the response variable, diagnosis.

RF using under-sampling

```
cv_us <- trainControl(method = "repeatedcv",
                      number = 10,
                      repeats = 10,
                      verboseIter = FALSE,
                      sampling="down")

set.seed(777)
modelrf_us <- caret::train(diagnosis ~ .,
                           data = train,
                           method = "rf",
                           trControl = cv_us)

final_us <- data.frame(actual = test$diagnosis,
                       predict(modelrf_us, newdata = test, type = "prob"))
final_us$predict <- ifelse(final_us$X0 > 0.5, 0, 1)
final_us$predict <- as.factor(final_us$predict)
cmRF_under <- confusionMatrix(final_us$predict, test$diagnosis)
```

```
## Warning in confusionMatrix.default(final_us$predict, test$diagnosis): Levels are
## not in the same order for reference and data. Refactoring data to match.
```

```
cmRF_under
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   0
##          1   9  16
##          0   4 101
##
##                Accuracy : 0.8462
##                  95% CI : (0.7724, 0.9034)
##     No Information Rate : 0.9
##     P-Value [Acc > NIR] : 0.98107
##
##                   Kappa : 0.3939
##
##  Mcnemar's Test P-Value : 0.01391
```

```
##
##             Sensitivity : 0.69231
##             Specificity : 0.86325
##          Pos Pred Value : 0.36000
##          Neg Pred Value : 0.96190
##              Prevalence : 0.10000
##          Detection Rate : 0.06923
##    Detection Prevalence : 0.19231
##       Balanced Accuracy : 0.77778
##
##        'Positive' Class : 1
##
```

```
## Printing cmRF_under shows true cancer cases misclassified 4/13 times (85% total accuracy)
```

RF using oversampling

```
cv_os <- trainControl(method = "repeatedcv",
                      number = 10,
                      repeats = 10,
                      verboseIter = FALSE,
                      sampling="up")

set.seed(777)
modelrf_os <- caret::train(diagnosis ~ .,
                           data = train,
                           method = "rf",
                           trControl = cv_os)

final_os <- data.frame(actual = test$diagnosis,
                       predict(modelrf_os, newdata = test, type = "prob"))
final_os$predict <- ifelse(final_os$X0 > 0.5, 0, 1)
final_os$predict <- as.factor(final_os$predict)
cmRF_over <- confusionMatrix(final_os$predict, test$diagnosis)
```

```
## Warning in confusionMatrix.default(final_os$predict, test$diagnosis): Levels are
## not in the same order for reference and data. Refactoring data to match.
```

```
cmRF_over
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   0
##          1   5   7
##          0   8 110
##
##                Accuracy : 0.8846
##                  95% CI : (0.8168, 0.934)
##     No Information Rate : 0.9
##     P-Value [Acc > NIR] : 0.7732
##
##                   Kappa : 0.3363
##
##  Mcnemar's Test P-Value : 1.0000
```

```
##
##             Sensitivity : 0.38462
##             Specificity : 0.94017
##          Pos Pred Value : 0.41667
##          Neg Pred Value : 0.93220
##              Prevalence : 0.10000
##          Detection Rate : 0.03846
##    Detection Prevalence : 0.09231
##       Balanced Accuracy : 0.66239
##
##        'Positive' Class : 1
##
```

```
## Printing cmRF_over shows true cancer cases misclassified 8/13 times (88% total accuracy)
```

RF using SMOTE

```
cv_smote <- trainControl(method = "repeatedcv",
                      number = 10,
                      repeats = 10,
                      verboseIter = FALSE,
                      sampling="smote")

set.seed(777)
modelrf_smote <- caret::train(diagnosis ~ .,
                           data = train,
                           method = "rf",
                           trControl = cv_smote)

final_smote <- data.frame(actual = test$diagnosis,
                          predict(modelrf_smote, newdata = test, type = "prob"))

final_smote$predict <- ifelse(final_smote$X0 > 0.5, 0, 1)
final_smote$predict <- as.factor(final_smote$predict)
cmRF_smote <- confusionMatrix(final_smote$predict, test$diagnosis)
```

```
## Warning in confusionMatrix.default(final_smote$predict, test$diagnosis): Levels
## are not in the same order for reference and data. Refactoring data to match.
```

```
cmRF_smote
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   0
##          1  10  13
##          0   3 104
##
##                Accuracy : 0.8769
##                  95% CI : (0.8078, 0.928)
##     No Information Rate : 0.9
##     P-Value [Acc > NIR] : 0.84704
##
##                   Kappa : 0.4904
##
```

```
##  Mcnemar's Test P-Value : 0.02445
##
##             Sensitivity : 0.76923
##             Specificity : 0.88889
##          Pos Pred Value : 0.43478
##          Neg Pred Value : 0.97196
##              Prevalence : 0.10000
##          Detection Rate : 0.07692
##    Detection Prevalence : 0.17692
##       Balanced Accuracy : 0.82906
##
##        'Positive' Class : 1
##
```

```
## Printing cmRF_SMOTE shows true cancer cases misclassified 3/13 times (88% total accuracy)
```

RF using ROSE

```
cv_rose <- trainControl(method = "repeatedcv",
                        number = 10,
                        repeats = 10,
                        verboseIter = FALSE,
                        sampling="rose")

set.seed(777)
modelrf_rose <- caret::train(diagnosis ~ .,
                             data = train,
                             method = "rf",
                             trControl = cv_rose)

final_rose <- data.frame(actual = test$diagnosis,
                         predict(modelrf_rose, newdata = test, type = "prob"))
final_rose$predict <- ifelse(final_rose$X0 > 0.5, 0, 1)
final_rose$predict <- as.factor(final_rose$predict)
cmRF_rose <- confusionMatrix(final_rose$predict, test$diagnosis)
```

```
## Warning in confusionMatrix.default(final_rose$predict, test$diagnosis): Levels
## are not in the same order for reference and data. Refactoring data to match.
```

```
cmRF_rose
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  0
##          1  8 19
##          0  5 98
##
##                Accuracy : 0.8154
##                  95% CI : (0.7379, 0.878)
##     No Information Rate : 0.9
##     P-Value [Acc > NIR] : 0.998931
##
##                   Kappa : 0.3064
##
```

```
##  Mcnemar's Test P-Value : 0.007963
##
##             Sensitivity : 0.61538
##             Specificity : 0.83761
##          Pos Pred Value : 0.29630
##          Neg Pred Value : 0.95146
##              Prevalence : 0.10000
##          Detection Rate : 0.06154
##    Detection Prevalence : 0.20769
##       Balanced Accuracy : 0.72650
##
##        'Positive' Class : 1
##
```

```
## Printing cmRF_ROSE shows true cancer cases misclassified 5/13 times (82% accuracy)
```

Now, for comparison, we will make the above models, but using a gradient boosting machine (GBM) algorithm, to see how the random forest (RF) model with highest sensitivity compares to the GBM model with the highest sensitivity. Once all models are trained and tested, we will compare their performances based on confusion matrix metrics, in a plot.

GBM using under-sampling

```
# Using cv_us as previously defined
set.seed(777)
modelgbm_us <- caret::train(diagnosis ~ .,
                            data = train,
                            method = "gbm",
                            verbose = FALSE,
                            trControl = cv_us)

final_usGBM <- data.frame(actual = test$diagnosis,
                          predict(modelgbm_us, newdata = test, type = "prob"))
final_usGBM$predict <- ifelse(final_usGBM$X0 > 0.5, 0, 1)
final_usGBM$predict <- as.factor(final_usGBM$predict)
cmGBM_under <- confusionMatrix(final_usGBM$predict, test$diagnosis)
```

```
## Warning in confusionMatrix.default(final_usGBM$predict, test$diagnosis): Levels
## are not in the same order for reference and data. Refactoring data to match.
```

```
cmGBM_under
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  0
##          1 10 20
##          0  3 97
##
##                Accuracy : 0.8231
```

```
##                 95% CI : (0.7465, 0.8844)
##     No Information Rate : 0.9
##     P-Value [Acc > NIR] : 0.9976476
##
##                   Kappa : 0.3784
##
##  Mcnemar's Test P-Value : 0.0008492
##
##             Sensitivity : 0.76923
##             Specificity : 0.82906
##          Pos Pred Value : 0.33333
##          Neg Pred Value : 0.97000
##              Prevalence : 0.10000
##          Detection Rate : 0.07692
##    Detection Prevalence : 0.23077
##       Balanced Accuracy : 0.79915
##
##        'Positive' Class : 1
##
```

```
## Printing cmGBM_under shows true cancer cases misclassified 3/13 times (82% total accuracy)
```

GBM using oversampling

```
# Using cv_os as previously defined
set.seed(777)
modelgbm_os <- caret::train(diagnosis ~ .,
                            data = train,
                            method = "gbm",
                            verbose = FALSE,
                            trControl = cv_os)

final_osGBM <- data.frame(actual = test$diagnosis,
                          predict(modelgbm_os, newdata = test, type = "prob"))
final_osGBM$predict <- ifelse(final_osGBM$X0 > 0.5, 0, 1)
final_osGBM$predict <- as.factor(final_osGBM$predict)
cmGBM_over <- confusionMatrix(final_osGBM$predict, test$diagnosis)
```

```
## Warning in confusionMatrix.default(final_osGBM$predict, test$diagnosis): Levels
## are not in the same order for reference and data. Refactoring data to match.
```

```
cmGBM_over
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   0
##          1   7   8
##          0   6 109
##
##                Accuracy : 0.8923
##                  95% CI : (0.8259, 0.9399)
##     No Information Rate : 0.9
##     P-Value [Acc > NIR] : 0.6807
##
```

```
##                  Kappa : 0.44
##
##  Mcnemar's Test P-Value : 0.7893
##
##             Sensitivity : 0.53846
##             Specificity : 0.93162
##          Pos Pred Value : 0.46667
##          Neg Pred Value : 0.94783
##              Prevalence : 0.10000
##          Detection Rate : 0.05385
##    Detection Prevalence : 0.11538
##       Balanced Accuracy : 0.73504
##
##        'Positive' Class : 1
##
```

```
## Printing cmGBM_over shows true cancer cases misclassified 6/13 times (89% total accuracy)
```

GBM using SMOTE

```
# Using cv_smote as previously defined
set.seed(777)
modelgbm_smote <- caret::train(diagnosis ~ .,
                           data = train,
                           method = "gbm",
                           verbose = FALSE,
                           trControl = cv_smote)

final_smoteGBM <- data.frame(actual = test$diagnosis,
                         predict(modelgbm_smote, newdata = test, type = "prob"))
final_smoteGBM$predict <- ifelse(final_smoteGBM$X0 > 0.5, 0, 1)
final_smoteGBM$predict <- as.factor(final_smoteGBM$predict)
cmGBM_smote <- confusionMatrix(final_smoteGBM$predict, test$diagnosis)
```

```
## Warning in confusionMatrix.default(final_smoteGBM$predict, test$diagnosis):
## Levels are not in the same order for reference and data. Refactoring data to
## match.
```

```
cmGBM_smote
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   0
##          1   9  16
##          0   4 101
##
##                Accuracy : 0.8462
##                  95% CI : (0.7724, 0.9034)
##     No Information Rate : 0.9
##     P-Value [Acc > NIR] : 0.98107
##
##                   Kappa : 0.3939
##
##  Mcnemar's Test P-Value : 0.01391
```

```
##
##            Sensitivity : 0.69231
##            Specificity : 0.86325
##         Pos Pred Value : 0.36000
##         Neg Pred Value : 0.96190
##             Prevalence : 0.10000
##         Detection Rate : 0.06923
##   Detection Prevalence : 0.19231
##      Balanced Accuracy : 0.77778
##
##       'Positive' Class : 1
##
```

```
## Printing cmGBM_smote shows true cancer cases misclassified 4/13 times (85% total accuracy)
```

GBM using ROSE

```
# Using cv_rose as previously defined
set.seed(777)
modelgbm_rose <- caret::train(diagnosis ~ .,
                              data = train,
                              method = "gbm",
                              verbose = FALSE,
                              trControl = cv_rose)

final_roseGBM <- data.frame(actual = test$diagnosis,
                            predict(modelgbm_rose, newdata = test, type = "prob"))
final_roseGBM$predict <- ifelse(final_roseGBM$X0 > 0.5, 0, 1)
final_roseGBM$predict <- as.factor(final_roseGBM$predict)
cmGBM_rose <- confusionMatrix(final_roseGBM$predict, test$diagnosis)
```

```
## Warning in confusionMatrix.default(final_roseGBM$predict, test$diagnosis):
## Levels are not in the same order for reference and data. Refactoring data to
## match.
```

```
cmGBM_rose
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   0
##          1   9  15
##          0   4 102
##
##                Accuracy : 0.8538
##                  95% CI : (0.7812, 0.9097)
##     No Information Rate : 0.9
##     P-Value [Acc > NIR] : 0.96557
##
##                   Kappa : 0.4099
##
##  Mcnemar's Test P-Value : 0.02178
##
##             Sensitivity : 0.69231
##             Specificity : 0.87179
```

```
##          Pos Pred Value : 0.37500
##          Neg Pred Value : 0.96226
##              Prevalence : 0.10000
##          Detection Rate : 0.06923
##    Detection Prevalence : 0.18462
##       Balanced Accuracy : 0.78205
##
##        'Positive' Class : 1
##
```

```
## Printing cmGBM_rose shows true cancer cases misclassified 4/13 times (85% total accuracy)
```

GBM model without stratified train/test split, and without sampling methods

(for comparison's sake)

```
#using cv_client as defined for the very first model
set.seed(777)
modelgbm_client <- caret::train(diagnosis ~ .,
                            data = train_client,
                            method = "gbm",
                            verbose = FALSE,
                            trControl = cv_client)

final_clientGBM <- data.frame(actual = test_client$diagnosis,
                          predict(modelgbm_client, newdata = test_client, type = "prob"))
final_clientGBM$predict <- ifelse(final_clientGBM$X0 > 0.5, 0, 1)
final_clientGBM$predict <- as.factor(final_clientGBM$predict)
cmGBM_imbalanced <- confusionMatrix(final_clientGBM$predict, test_client$diagnosis)
```

```
## Warning in confusionMatrix.default(final_clientGBM$predict,
## test_client$diagnosis): Levels are not in the same order for reference and data.
## Refactoring data to match.
```

```
cmGBM_imbalanced
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   1   0
##          1   5   5
##          0   8 113
##
##                Accuracy : 0.9008
##                  95% CI : (0.8363, 0.9461)
##     No Information Rate : 0.9008
##     P-Value [Acc > NIR] : 0.5731
##
##                   Kappa : 0.3814
##
##  Mcnemar's Test P-Value : 0.5791
##
##            Sensitivity : 0.38462
##            Specificity : 0.95763
##          Pos Pred Value : 0.50000
```

```
##          Neg Pred Value : 0.93388
##              Prevalence : 0.09924
##          Detection Rate : 0.03817
##    Detection Prevalence : 0.07634
##       Balanced Accuracy : 0.67112
##
##        'Positive' Class : 1
##
```

```
## Printing cmGBM_noSampling shows true cancer cases misclassified 8/13 times (90% total accuracy)
```

Comparing confusion matrix metrics of each of the 10 models (the focus here is on sensitivity)

```
#RF
models <- list(imbalancedClient = modelrf_client,
               under = modelrf_us,
               over = modelrf_os,
               smote = modelrf_smote,
               rose = modelrf_rose)

comparison1 <- data.frame(model = names(models))

for (name in names(models)) {
  model <- get(paste0("cmRF_", name))

  comparison1[comparison1$model == name, "Sensitivity"] <- model$byClass[["Sensitivity"]]
  comparison1[comparison1$model == name, "Specificity"] <- model$byClass[["Specificity"]]
}

comparison1 %>%
  gather(x, y, Sensitivity:Specificity)
```

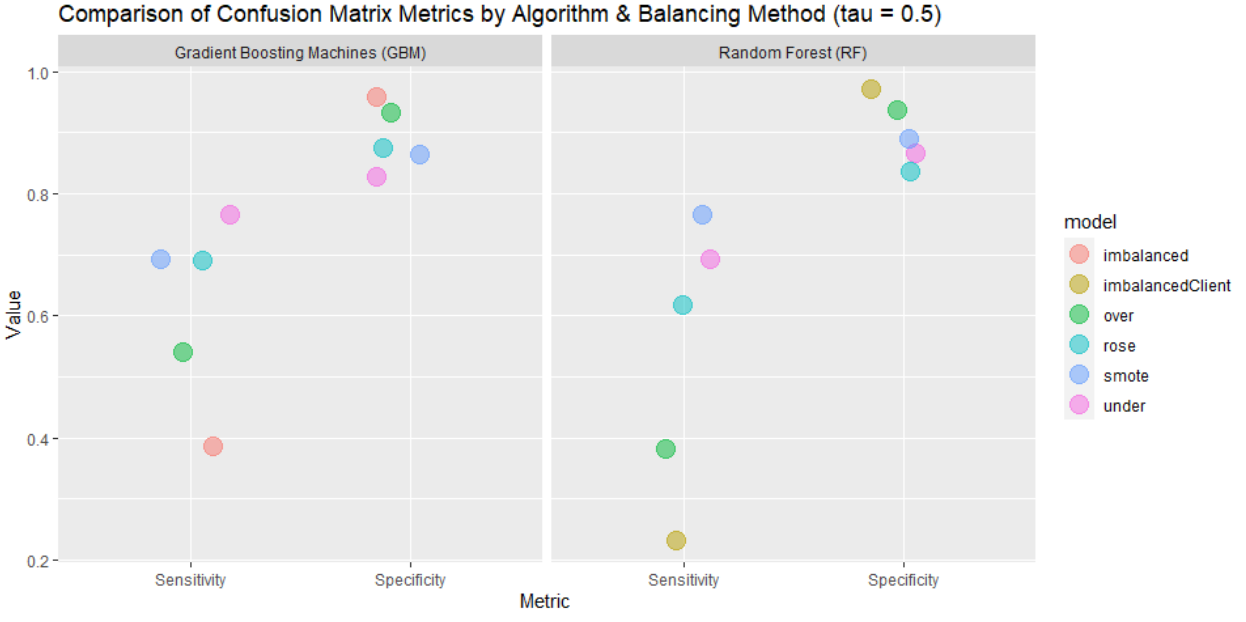
```
##               model           x         y
## 1  imbalancedClient Sensitivity 0.2307692
## 2             under Sensitivity 0.6923077
## 3              over Sensitivity 0.3846154
## 4             smote Sensitivity 0.7692308
## 5              rose Sensitivity 0.6153846
## 6  imbalancedClient Specificity 0.9745763
## 7             under Specificity 0.8632479
## 8              over Specificity 0.9401709
## 9             smote Specificity 0.8888889
## 10             rose Specificity 0.8376068
```

```
#GBM
models <- list(imbalanced = modelgbm_client,
               under = modelgbm_us,
               over = modelgbm_os,
               smote = modelgbm_smote,
               rose = modelgbm_rose)

comparison2 <- data.frame(model = names(models))

for (name in names(models)) {
```

```
  model <- get(paste0("cmGBM_", name))

  comparison2[comparison2$model == name, "Sensitivity"] <- model$byClass[["Sensitivity"]]
  comparison2[comparison2$model == name, "Specificity"] <- model$byClass[["Specificity"]]
}

comparison2 %>%
  gather(x, y, Sensitivity:Specificity)
```

```
##         model           x         y
## 1  imbalanced Sensitivity 0.3846154
## 2       under Sensitivity 0.7692308
## 3        over Sensitivity 0.5384615
## 4       smote Sensitivity 0.6923077
## 5        rose Sensitivity 0.6923077
## 6  imbalanced Specificity 0.9576271
## 7       under Specificity 0.8290598
## 8        over Specificity 0.9316239
## 9       smote Specificity 0.8632479
## 10       rose Specificity 0.8717949
```

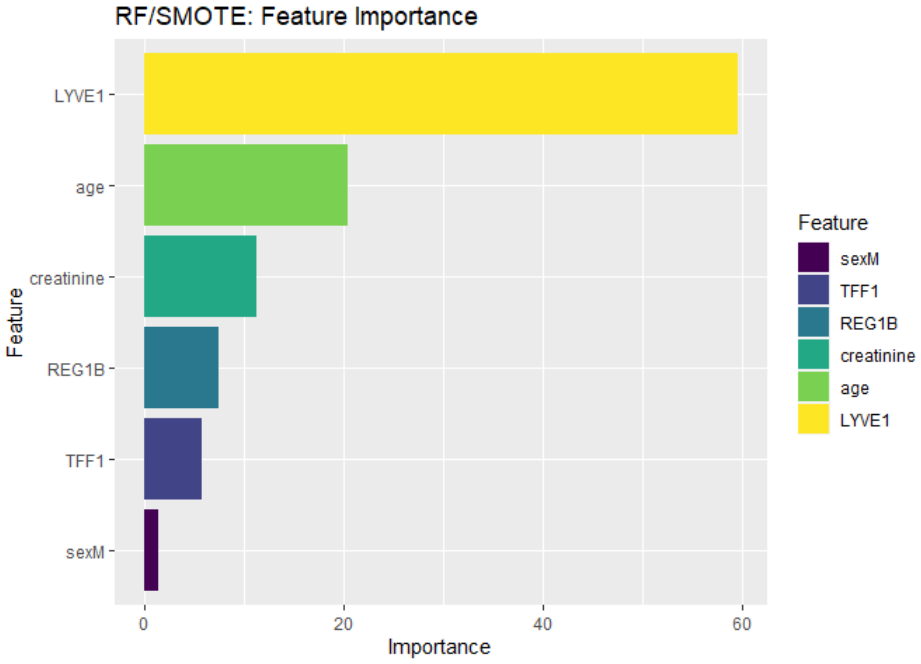
```
#the plot
comparison1 <- cbind(comparison1,"Random Forest (RF)")
colnames(comparison1)[4] <- 'algo'
comparison2 <- cbind(comparison2,"Gradient Boosting Machines (GBM)")
colnames(comparison2)[4] <- 'algo'
both <- rbind(comparison1,comparison2)

both %>%
  gather(x, y, Sensitivity:Specificity) %>%
ggplot(aes(x=x,y=y, color = model)) +
  geom_jitter(width = 0.2, alpha = 0.5, size = 5) +
  facet_wrap(~algo) +
  labs(title="Comparison of Confusion Matrix Metrics by Algorithm & Balancing Method (tau = 0.5)") +
  labs(x="Metric",y="Value")
```



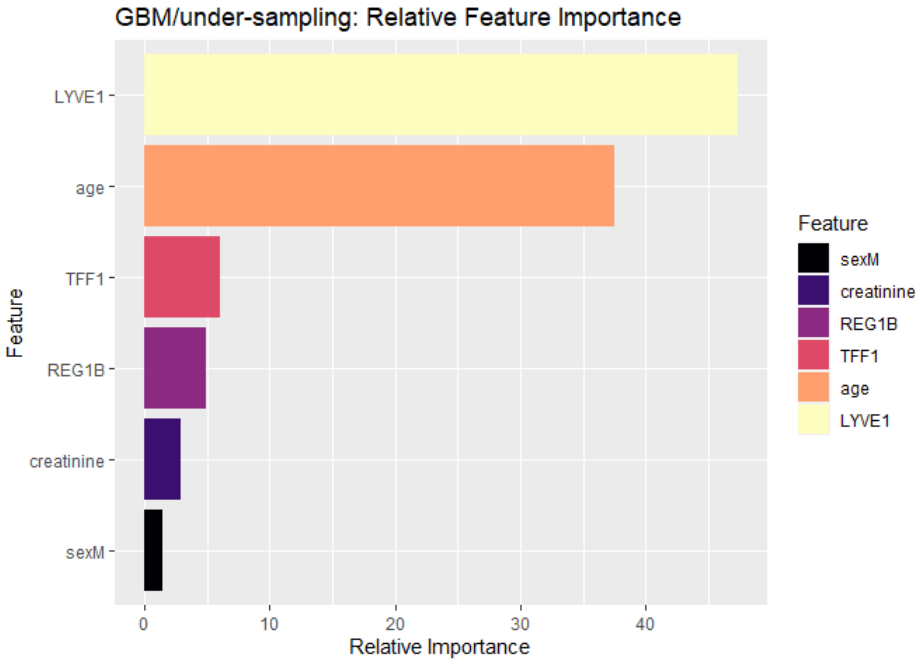
We see that SMOTE yields the highest sensitivity for the RF model, and under-sampling yields the highest sensitivity for the GBM model, but they appear to be tied in sensitivity with RF/SMOTE having slightly higher specificity than GBM/under-sampling. Before we try tuning these 2 models, we can get feature importance plots, as these are invariable to tuning).

Check variable importance for the best 2 models

```
# RF/SMOTE
VI <- varImp(modelrf_smote, scale=FALSE)
VI <- VI$importance
VI$feat <- row.names(VI)
VI <- transform(VI, feat = reorder(feat, Overall))
ggplot2::ggplot(VI, aes(Overall, feat, fill = feat)) +
  geom_col(aes()) +
  labs(title="RF/SMOTE: Feature Importance") +
  labs(x="Importance",y="Feature") +
    scale_fill_viridis_d(option="viridis", name = "Feature")
```




```
# GBM/Under
RI <- summary(modelgbm_us, plot = FALSE)
RI <- transform(RI, var = reorder(var, rel.inf))
ggplot2::ggplot(RI, aes(rel.inf, var, fill = var)) +
  geom_col(aes()) +
  labs(title="GBM/under-sampling: Relative Feature Importance") +
  labs(x="Relative Importance",y="Feature") +
    scale_fill_viridis_d(option="magma", name = "Feature")
```

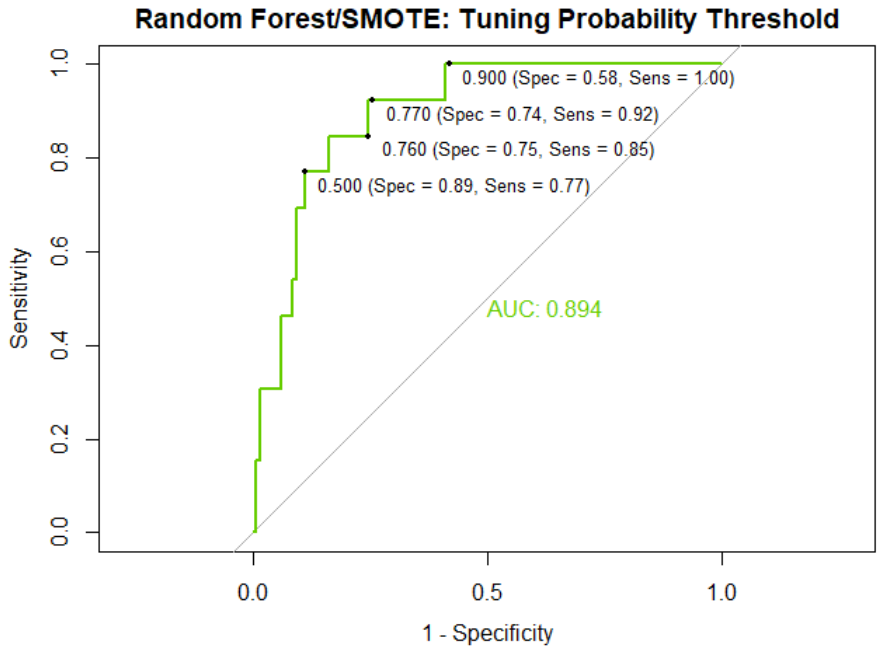


Find optimal probabiltity threshold

```
## Get the ROC curve for Random Forest/SMOTE model
roc_rf <- roc(test$diagnosis,
              predict(modelrf_smote, test, type = "prob")[,2],
              levels = rev(levels(test$diagnosis)))
```

```
## Setting direction: controls > cases
```

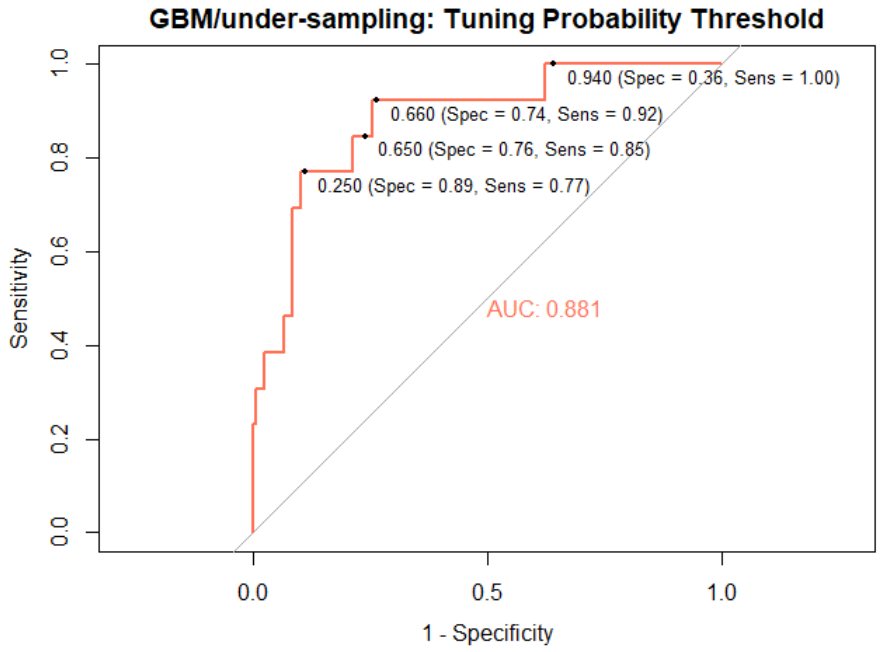
```
#roc_rf


plot(roc_rf, print.thres = c(0.5, 0.76, 0.77, 0.9), type = "S", col = "chartreuse3",
     print.thres.pattern = "%.3f (Spec = %.2f, Sens = %.2f)",
     print.thres.cex = .8,
     print.auc = TRUE,
     legacy.axes = TRUE, main = "Random Forest/SMOTE: Tuning Probability Threshold")
```



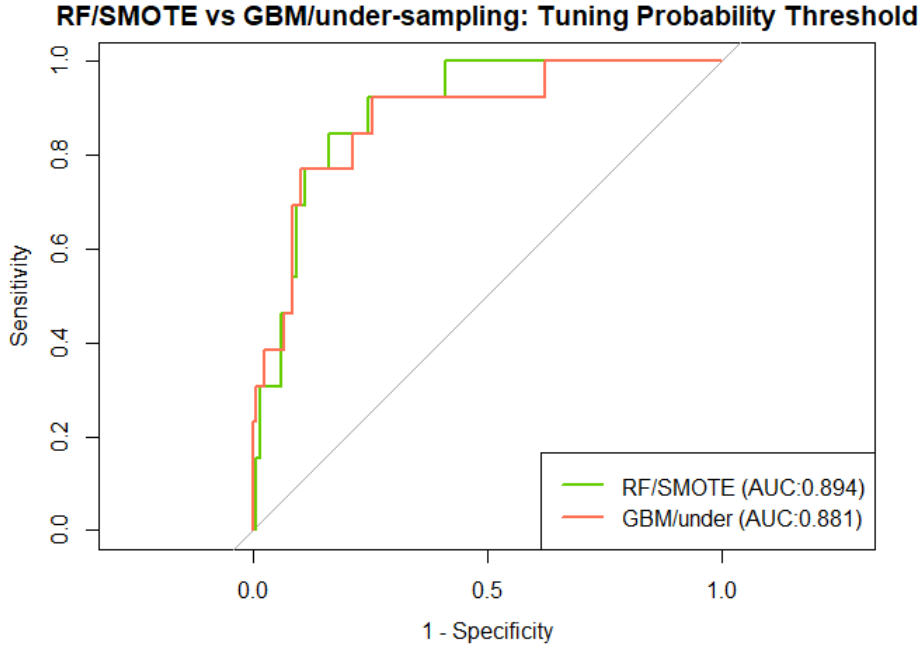

```
## Get the ROC curve for GBM/under-sampling model
roc_gbm <- roc(test$diagnosis,
            predict(modelgbm_us, test, type = "prob")[,2],
            levels = rev(levels(test$diagnosis)))
```

```
## Setting direction: controls > cases
```

```
#roc_gbm


plot(roc_gbm, print.thres = c(0.25, 0.65, 0.66, 0.94), type = "S", col = "coral1",
     print.thres.pattern = "%.3f (Spec = %.2f, Sens = %.2f)",
     print.thres.cex = .8,
     print.auc = TRUE,
     legacy.axes = TRUE, main = "GBM/under-sampling: Tuning Probability Threshold")
```




```
#plot both together
plot(roc_rf, type = "S", col = "chartreuse3",
     print.thres.pattern = "%.3f (Spec = %.2f, Sens = %.2f)",
     print.thres.cex = .8,
     legacy.axes = TRUE, main = "RF/SMOTE vs GBM/under-sampling: Tuning Probability Threshold")
lines(roc_gbm, type = "S", col = "coral1",
     print.thres.pattern = "%.3f (Spec = %.2f, Sens = %.2f)",
     print.thres.cex = .8,
     legacy.axes = TRUE, main="")
legend("bottomright", c("RF/SMOTE (AUC:0.894)","GBM/under (AUC:0.881)"),
       col=c("chartreuse3","coral1"), lty=c(1,1), lwd=c(2,2))
```



Make a table of RF/SMOTE thresholds and corresponding sensitivities & specificities

```
TH <- roc_rf$thresholds[22:42]
SN <- roc_rf$sensitivities[22:42]
SP <- roc_rf$specificities[22:42]

df <- round(cbind(TH,SN,1-SP),3)

kable(df, col.names = c("Threshold (tau)","Sensitivity (Prop. of true +)",
                        "1 - Specificity (Prop. of false +)"),
      caption = "RF/SMOTE:  Comparing Probability Thresholds")
```

RF/SMOTE: Comparing Probability Thresholds

Threshold (tau)	Sensitivity (Prop. of true +)	1 - Specificity (Prop. of false +)
0.913	1.000	0.436
0.909	1.000	0.427
0.903	1.000	0.419
0.898	1.000	0.410
0.895	0.923	0.402
0.892	0.923	0.393
0.887	0.923	0.385
0.881	0.923	0.368
0.877	0.923	0.359
0.875	0.923	0.350
0.873	0.923	0.342
0.871	0.923	0.316
0.869	0.923	0.299
0.859	0.923	0.291
0.845	0.923	0.282
0.832	0.923	0.274
0.810	0.923	0.265
0.782	0.923	0.256
0.767	0.923	0.248
0.763	0.846	0.248
0.750	0.846	0.231

Comparing RF/SMOTE sensitivities: tau = 0.5 vs. tau = 0.77

```
#using CV_smote, modelrf_smote, and final_smote as previously defined

final_smote$predictTuned <- ifelse(final_smote$X0 > 0.77, 0, 1)
```

```
final_smote$predictTuned <- as.factor(final_smote$predictTuned)
cmRF_smoteTuned <- confusionMatrix(final_smote$predictTuned, test$diagnosis)
```

```
## Warning in confusionMatrix.default(final_smote$predictTuned, test$diagnosis):
## Levels are not in the same order for reference and data. Refactoring data to
## match.
```

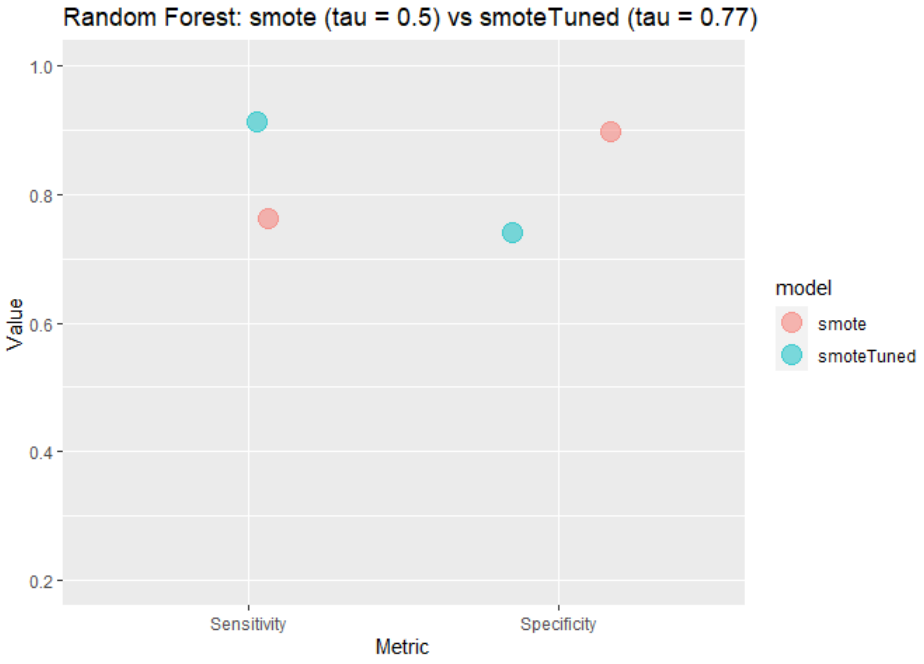
```
cmRF_smoteTuned
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  0
##          1 12 30
##          0  1 87
##
##                Accuracy : 0.7615
##                  95% CI : (0.6789, 0.8319)
##     No Information Rate : 0.9
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3348
##
##  Mcnemar's Test P-Value : 4.932e-07
##
##             Sensitivity : 0.92308
##             Specificity : 0.74359
##          Pos Pred Value : 0.28571
##          Neg Pred Value : 0.98864
##              Prevalence : 0.10000
##          Detection Rate : 0.09231
##    Detection Prevalence : 0.32308
##       Balanced Accuracy : 0.83333
##
##        'Positive' Class : 1
##
```

Plot confusion matrix metrics comparing RF/SMOTE using tau = 0.50 vs tau = 0.77

```
# label models
models <- list(smote = modelrf_smote,
               smoteTuned = modelrf_smote)

# data manipulation to prep for plot
comparison <- data.frame(model = names(models))

for (name in names(models)) {
  model <- get(paste0("cmRF_", name))

  comparison[comparison$model == name, "Sensitivity"] <- model$byClass[["Sensitivity"]]
  comparison[comparison$model == name, "Specificity"] <- model$byClass[["Specificity"]]
}

# make the plot
```

```
comparison %>%
  gather(x, y, Sensitivity:Specificity) %>%
  ggplot(aes(x = x, y = y, color = model)) +
  geom_jitter(width = 0.2, alpha = 0.5, size = 5) +
  labs(title="Random Forest: smote (tau = 0.5) vs smoteTuned (tau = 0.77)") +
  labs(x="Metric",y="Value") +
  ylim(0.2, 1.0)
```



Choosing the RF/SMOTE classifier procedure followed by tuning its threshold to tau = 0.77 accomplishes the stated goal of reaching true positive rate (sensitivity) of at least .90 while keeping false positive rate (1-specificity) from surpassing 0.25. The GBM/under-sampling classifier procedure can accomplish the same goal using tau = 0.66, however the RF/SMOTE classifier is recommended as it not only has a slightly higher overall predicitive performance (AUC), but it also shows a slightly lower rate of false positives at a sensitivity of 0.92, and a lower rate of false positives for sensitivities of 0.78 to 1.0, making it the much preferred procedure even if goal limits change moderately in either direction.