Title: Predicting Breast Cancer Survivability

Authors: Vinu Baburaj, Matthew Greene, Hemashree Kilari, Atharva Abhijit Kulkarni, Shashank Bettada Sathya Thirtha

Summary:

Background and motivation: Breast cancer is the most prevalent form of cancer, which is also responsible for the highest number of cancer-related deaths among women. The American Cancer Society estimates that about 288,000 women worldwide will be diagnosed with breast cancer in 2022. Cancers are associated with genetic abnormalities. Gene expression measures the level of gene activity in a tissue and gives information about its complex activities. Comparing the genes expressed in normal and diseased tissue can bring better insights about the cancer prognosis and outcomes.

Project Goal: Our goal is to analyze the attributes in our dataset and find ones that display a relationship with the best chances of surviving a breast cancer diagnosis.

Description of work and results: We built four machine learning models to predict the likeliness of survival, based on the identified critical attributes. We then compared the results of each of our four models for the best fit. Our best model and results were using Random Forests with variable hyper tuning which achieved an accuracy of 72%.

Dataset Description: The dataset for this analysis and prediction is called Breast Cancer Gene Expression Profiles (METABRIC) which contains sequencing data of 1,980 primary breast cancer samples with 693 attributes and is found on Kaggle. Of the 693 variables, 31 were deemed critical. These included basic patient information such as age at diagnosis, the type of breast cancer surgery if they had one, Boolean values for if they had treatments like chemo, radiation, or hormone therapy, as well as if the patient survived, how long they have survived post diagnosis and if they had passed, either from breast cancer or another reason. The other 662 variables are made up of genetic mutation information for each patient that can usually be used to predict whether a patient genetically predisposed to get breast cancer or other diseases. However, since all the patients in this study already have breast cancer, the results from testing those variables were not as significant as it would be given populations that include those that do not have a breast cancer diagnosis (Breast cancer gene expression profiles (METABRIC) 2019).

Related Work: This project was derived from a dataset collected by Professors Carlos Caldas and Sam Aparicio that started a series of published papers on breast cancer patients and their tumors beginning in 2016.

Methods:

EDA:

We first began by looking at the data that was missing and noticed that there was over 26% missing from the Tumor stage column. This posed a bit of a problem as tumor stage seemed like it would play a large role in whether breast cancer would kill someone or not. However, we discovered that Nottingham Prognostic Index (NPI) was also included in the dataset. NPI is equal to [0.2 * Size of the

tumor (*in centimeters*)] + number of positively infected lymph nodes + the grade of the tumor. The NPI calculation ends up with a scale from approximately 2 to 6, where 2 is the lower end of the scale and 6 is the larger end of the scale. This means that as the index increases, the larger the tumor and the shorter time a patient is expected to live. Exploring the dataset further, we noticed that the data was already in Codd Third Normalized Form and didn't need any further tidying or transformation.

SVM:

"Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate.

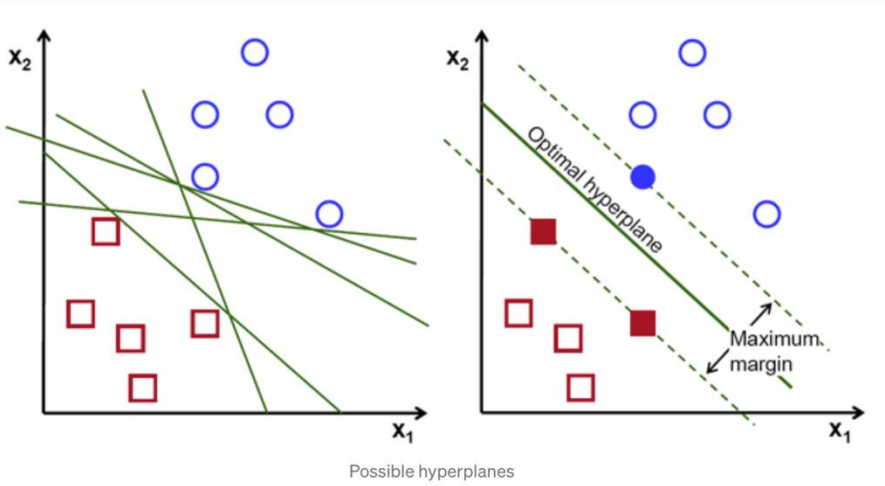

Figure 1: Hyperplane

SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane (MMH) that best divides the dataset into classes.

Logisitic Regression:

Logistic regression is a Supervised Machine Learning classification model which predicts values based on one or more variables and the answer is in a binary form, i.e., yes or no, 1 or 0. Logistic Regression applies sigmoid function to the input variables, and the graph produced after applying Logistic Regression is used to determine the value of the target variable. The Logistic Regression takes various assumptions into consideration like:

a. There is no relationship between the observation variables.
b. The target variable is binary.

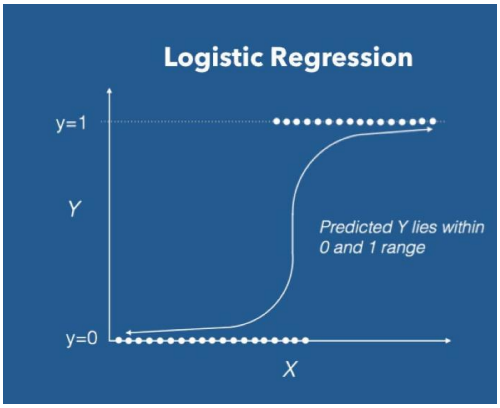

Figure 2 : Sigmoid graph for predicting the target variable value.

After performing pre-processing on the dataset, a total of 46 variables containing 1202 rows were taken into consideration for building the Logistic Regression Model. The dataset was then divided into 75% and 25% training set and testing set. To further enhance and improve the accuracy of the model, feature selection was performed. The feature selection process left us with the 15 most contributing features which were then fitted into the model and the accuracy was determined. (Kaiser, Predicting breast cancer using logistic regression 2020)

Naive Bayes:

Naïve Bayes is a probabilistic machine learning model based on Baye's theorem used for supervised classification. It works on the assumption that every feature is independent of each other. Naïve Bayes helps us to determine the probability of an outcome, given the probability of events that have taken place. We found naïve bayes to be a suitable classifier to this project as the different variables can have a conditional influence on the overall outcome of survivability.

We started with 46 critical variables with 1202 rows of observations. We then divided the observations into training and testing variables containing 75% and 25% values respectively. This left us with 901 observations for training and 301 observations for testing. To narrow down on the features of importance to our naïve bayes classifier, we performed forward feature selection based on model accuracy. This provided us with 14 features of importance like age at diagnosis, hormone therapy, mutation count, tumor size, cellularity, pm50 + claudin etc. We then fit these features into our gaussian naïve bayes classifier

Random Forests:

Random Forest is a supervised classification algorithm that makes use of decision trees. A decision tree is a series of yes/no questions that lead to the outcome in the form of our target variable. The finals outcome of the tree is represented as the leaf node. The Random Forest combines multiple decision trees to form what is called a decision forest to give an accurate prediction.

Results:

EDA:

During the EDA, we focused on three main questions to answer. Those questions were, what variables best predict whether a patient will survive? For all patients, how long do they live past their diagnosis? Finally, which treatments were most effective in extending the life of patients? These questions guided how we approached the EDA and drew conclusions to help build our models.

First, we plotted the different factors against overall survival and survival months. During this process, we noticed what looks like a very strong correlation between the age of diagnosis and whether someone survives their diagnosis. The median age of those that survived is approximately 55 at the time of diagnosis while the those that didn't were closer to 67 when diagnosed as seen in figure 1. There were no other variables that so clearly correlated to overall survival.

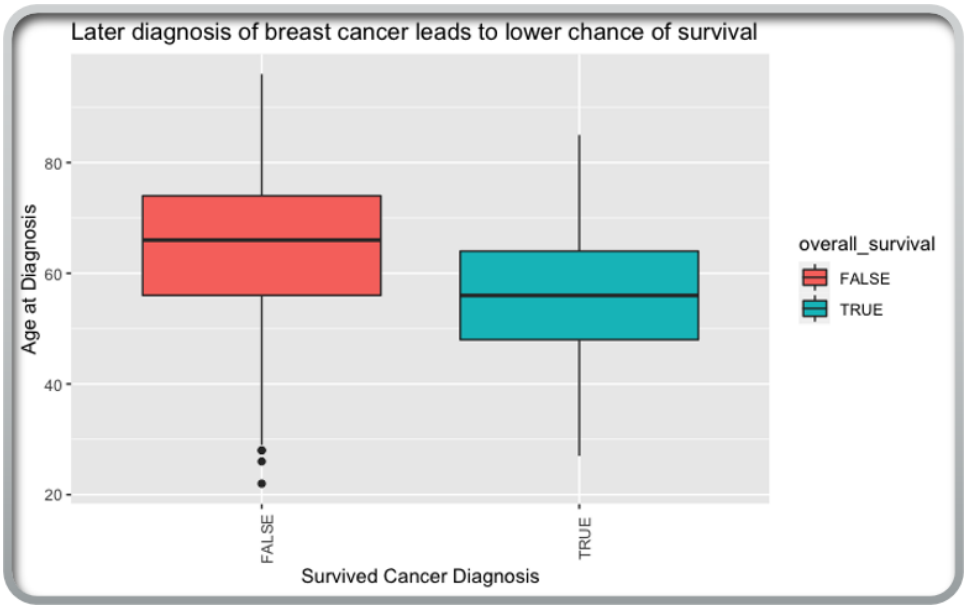

Figure 3: The age of Diagnosis vs Patient survival

Next, we explored variables versus how long a patient lived past their diagnosis. In doing so, we noticed another strong correlation between NPI score and overall survival based on if they were living had died from Breast Cancer or died from something else. We again noticed a very strong correlation between the higher NPI score and lower expected time to live among those that had died from Breast Cancer and those that had died from other causes. As seen in figure 2 on the left-hand side with the redline with red outline, as the NPI increases the months of survival sharply decrease under the Died of Disease category. The slope of the trend line in the Died of Other Causes graph (figure 2, red line with green outline) caused to conclude that there may be a connection between breast cancer and its associated treatments contributing to the death of patient.

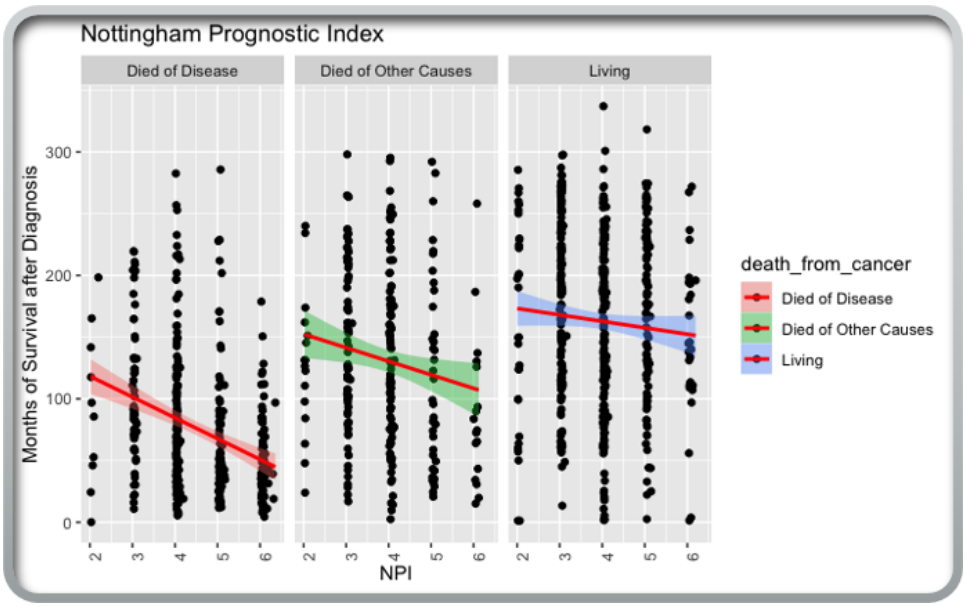

Figure 4: Nottingham Prognostic Index vs Months of Survival after Diagnosis

Finally, we explored which treatments and surgeries were most effective for extending the life of patients. In 99% of cases, patients had some combination of surgical (Mastectomy and Breast Conservation) as well as other chemo, radiation and hormone therapy. Chemotherapy and Mastectomy surgery seem to be the most effective treatments but are even more so when in combination of radiation treatment as seen in Figure 3.

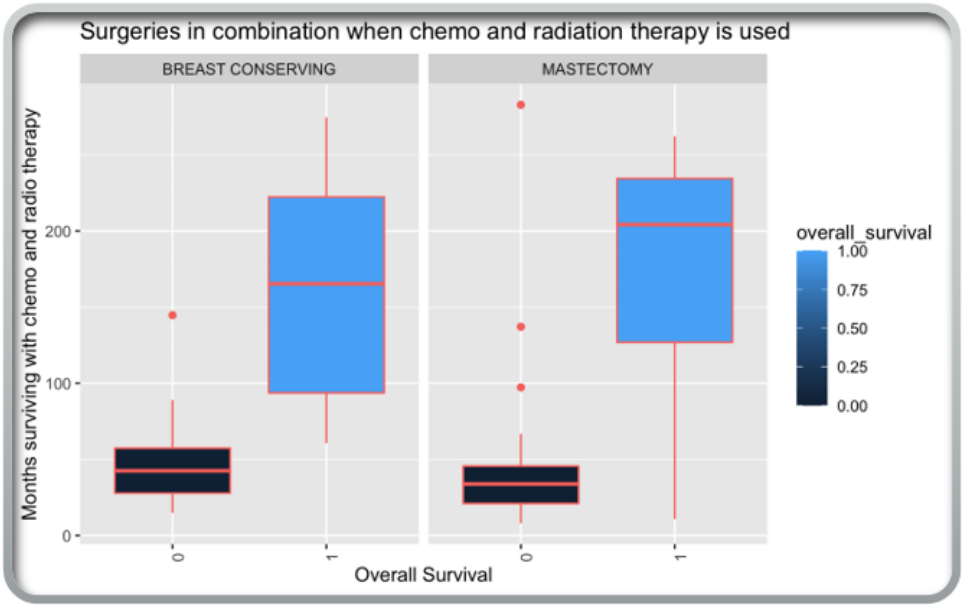

Figure 5: Months of survival vs overall survival when either surgical option is paired with chemotherapy and radiation

Modeling Results:

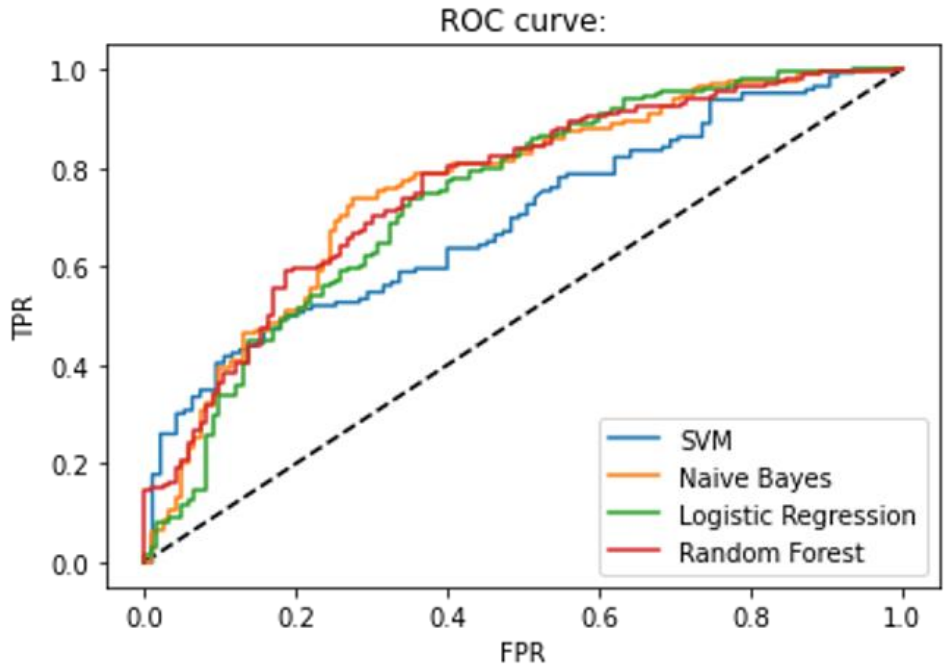

Figure 6: ROC curves of all 4 models

SVM:

There are 2 classes: "0" and "1" in our target variable. Using SVM after prediction, classifies the data into one of these classes. After pre-processing the data, we fitted the data using different SVM kernels like linear, sigmoid, rbf, and poly. In comparison to other kernels, the "linear" kernel provided the best accuracy of 69%. Then we plotted an ROC curve.

Logisitic Regression:

The accuracy achieved by the Logistic Regression model was 70%. While the accuracy is not up to the mark, achieving greater accuracy was difficult since there is a very little chance that all the observatory variables are not dependent on each other.

Naive Bayes:

We were able to achieve an accuracy of 70% with our naïve bayes classifier. While this accuracy does not seem too high, it was around the expected range considering that independence of all variables is difficult to achieve in real world applications.

Random Forests:

In order to reduce the risk of overfitting the data, we performed stepwise feature selection in the forward direction to get the set of variables that highly correlate with the target variable. To perform the stepwise feature selection, the estimator used was random forest itself. The feature selector returned 25 variables. Some of the key variables returned were 'age_at_diagnosis', 'lymph_nodes_examined_positive', 'mutation_count' and 'tumor_size'.

With the selected features only, the model was fitted with the train portion of the dataset. The model was tested on the test dataset and the resultant accuracy was 70%. However, for best results, it is imperative to tune the random forest estimator's hyper parameters. Post hyper parameter tuning, the model accuracy was improved to 72%.

Discussion:

Meaning: Our results demonstrate that it is very hard to predict whether someone will survive a breast cancer diagnosis. While we can predict with 72% confidence that someone will survive their diagnosis, further data collection and analysis are needed to improve the dataset and models.

Beneficiaries: Breast cancer patients and researchers are those that could benefit from our project. This analysis could be used and refined to build even more accurate models as well as data collection practices to allow for better prediction of breast cancer survival. Our results would be most useful to help determine treatment options for patients and serve as a starting point for patients and doctors to have conversations on treatment plans.

Future Work: Areas of improvement and further work include the data collection and recording methods, further sampling, and exploration of which genetic markers would be more useful for further analysis.

First, based on the combinations of treatments and surgeries as well as the medical advice of Dr. George Alverez (Alvarez, 2022), we believe that many of these patients underwent multiple rounds of treatments and that they may also have multiple diagnosis. Based on our interpretation of the data, it appears that patients may have had multiple diagnosis and gone into remission only for the cancer to return. The cancer may not have responded to single rounds of specific treatments, or the patient may have required multiple rounds of each. But we can't be certain because the treatments are only recorded as Booleans, meaning either they did or did not have the treatment. Having the number of rounds of each treatment, whether those treatments were effective, and how many times a patient was diagnosed along with the associated treatment for each diagnosis and the direct result would be helpful. Other pieces of data that would be helpful would be an estimation of if either the breast cancer or its associated treatment played a role in the death of patients who died of other causes. While the treatment and diagnosis may not be the direct reason for a patient passing, we find it reasonable to assume that the effects of treatment could have a severe impact on the function of other organs and bodily systems.

Second, further sampling would be helpful that includes patients that did not have breast cancer and what their genetic marker information was. While this would not directly affect our question and models, it may shed more light on how consequential some of the genetic markers are and if they may be better or earlier ways to detect breast cancer. Also, based on the prevalence of those markers, it may indicate how severely or when breast cancer would become diagnosable.

Finally, further work would include more exploration of the genetic marker data. A good course of action for this would be to include a medical researcher or doctor as either a consultant or fellow researcher.

Statement of contributions:

Vinu Baburaj: Performed the Naïve Bayes modeling and analysis.

Matthew Greene: Performed the exploratory data analysis, formatted and organized the project presentation and report.

Hemashree Kilari: Performed the SVM modeling and analysis.

Atharva Abhijit Kulkarni: Performed the Logistic Regression modeling and analysis.

Shashank Bettada Sathya Thirtha: Performed data preprocessing, Random Forest Modeling and organized the project proposal.

All team members participated in the writing, editing, and preparation of the project proposal, presentation, and report.

Bibliography

A. Mukherjee, R. Russell, S.-F. Chin, B. Liu, O. M. Rueda, H. R. Ali, G. Turashvili, B. Mahler-Araujo, I. O. Ellis, S. Aparicio, C. Caldas, and E. Provenzano, "Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort," *Nature News*, 07-Mar-2018. [Online]. Available: https://www.nature.com/articles/s41523-018-0056-8. [Accessed: 11-Dec-2022].

"Breast cancer gene expression profiles (METABRIC)," *Kaggle*, 2019. [Online]. Available: https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric. [Accessed: 11-Dec-2022].

"Breast cancer treatment (adult) (PDQ®)–patient version," *National Cancer Institute*, 2022. [Online]. Available: https://www.cancer.gov/types/breast/patient/breast-treatment-pdq. [Accessed: 11-Dec-2022].

E. A. Rakha, D. Soria, A. R. Green, C. Lemetre, D. G. Powe, C. C. Nolan, J. M. Garibaldi, G. Ball, and I. O. Ellis, "Nottingham Prognostic Index Plus (NPI+): A modern clinical decision making tool in breast cancer," *British journal of cancer*, 02-Apr-2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3974073/. [Accessed: 11-Dec-2022].

G. Alvarez, "Intracacies of Breast Cancer Diagnosis and Genetics," 28-Nov-2022.

M. Kaiser, "Predicting breast cancer using logistic regression," *Medium*, 17-Mar-2020. [Online]. Available: https://medium.com/swlh/predicting-breast-cancer-using-logistic-regression-3cbb796ab931. [Accessed: 11-Dec-2022].

"Stages of breast cancer: Understand breast cancer staging," *American Cancer Society*, 2021. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html. [Accessed: 11-Dec-2022].

Appendix:

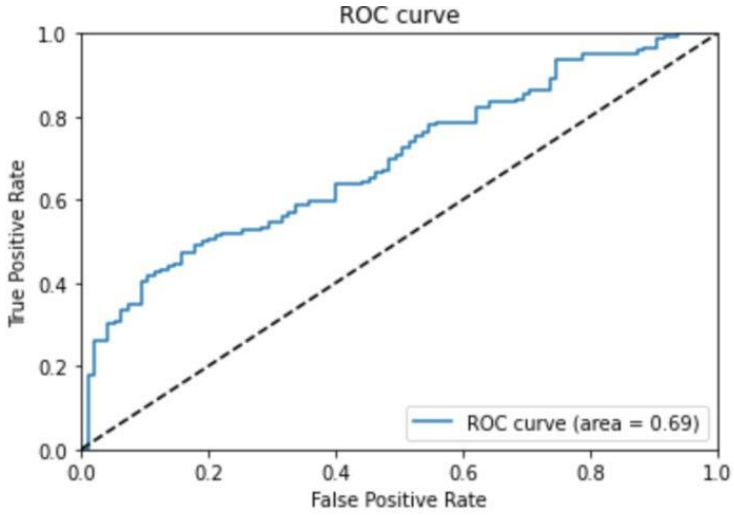

Figure 7: ROC Curve for Logistic Regression Model.

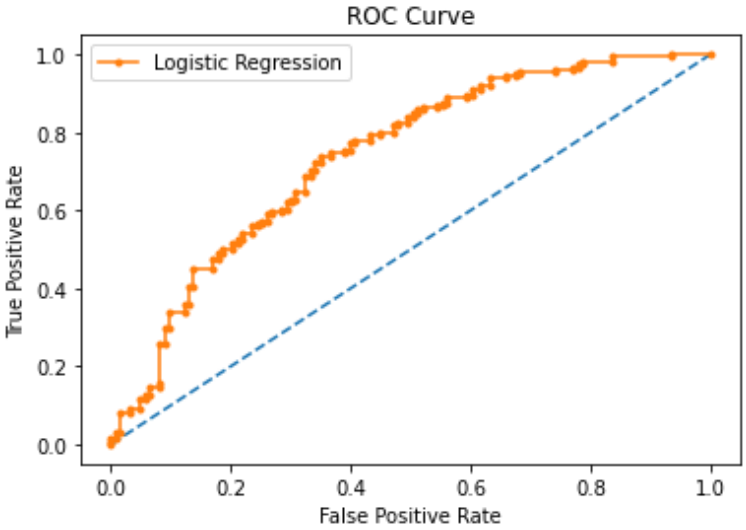

Figure 8: ROC Curve for Logistic Regression Model.

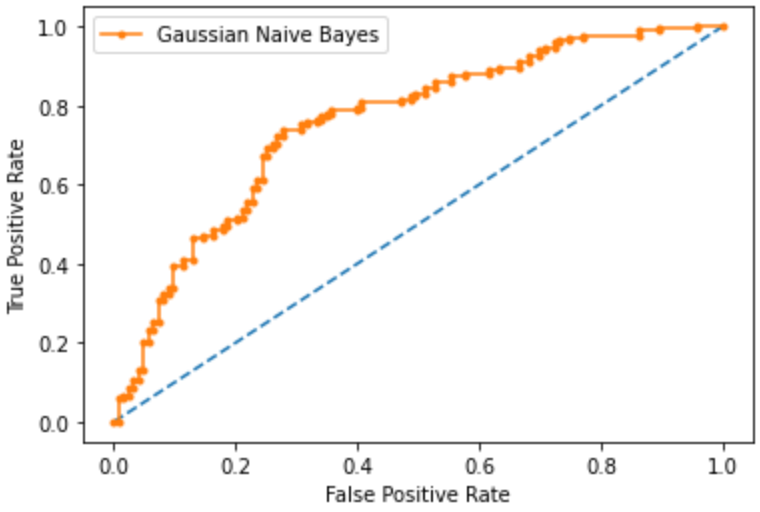

Figure 9: ROC curve for Naïve Bayes classifier model.

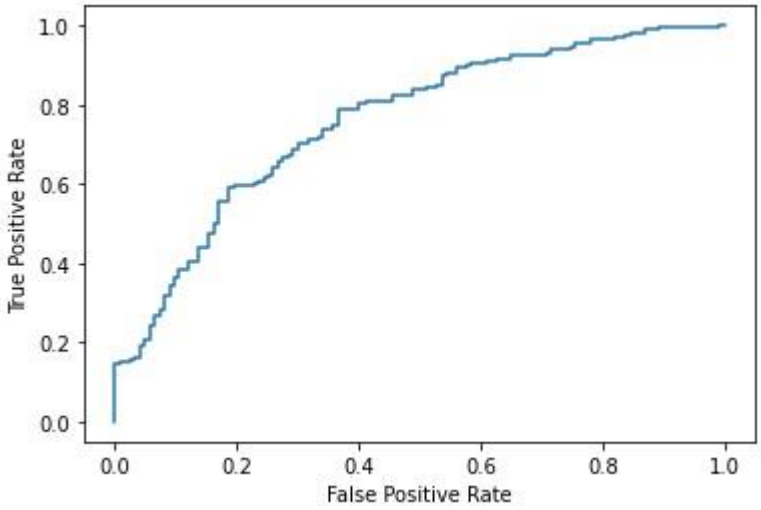

Figure 10: ROC curve for Random Forest classifier model.