Breast Cancer Prediction

Mehmet Erdoğdu

Linkedin:minanerdogdu@gmail.com
email:https://www.linkedin.com/in/mehmetierdogdu

Overview

If breast cancer is left untreated, the cancer spreads out to other parts of the body if it is a malignant cell growth

The average 10-year survival rate for women with invasive breast cancer is 84%. This year, an estimated 276,480 women in the United States will be diagnosed with invasive breast cancer, and It is estimated that 42,690 deaths (42,170 women and 520 men) from breast cancer will occur this year

Investigating the Data

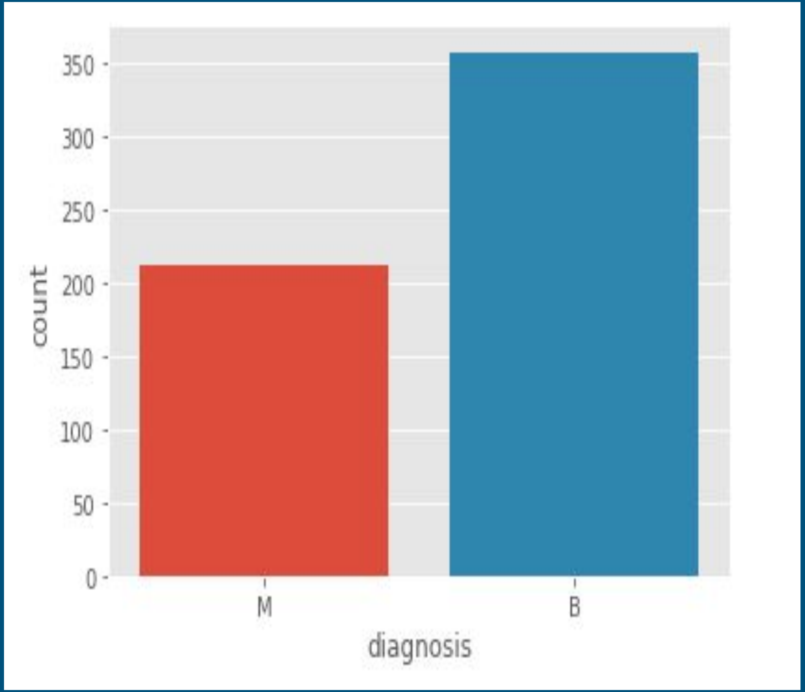
- The data set has been acquired through Kaggle website, which can be found at this link:
  Breast Cancer Wisconsin (Diagnostic) Data Set

- The dataset has 33 columns with 569 rows. There are ten computed real-valued features for each cell nucleus in the data set: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimensions.

- There are a total of 357 Benign and 212 malignant points in the data set

- Each feature has 3 corresponding columns - worst, mean, and standard error

- Reduced those features down to mitigate multicollinearity



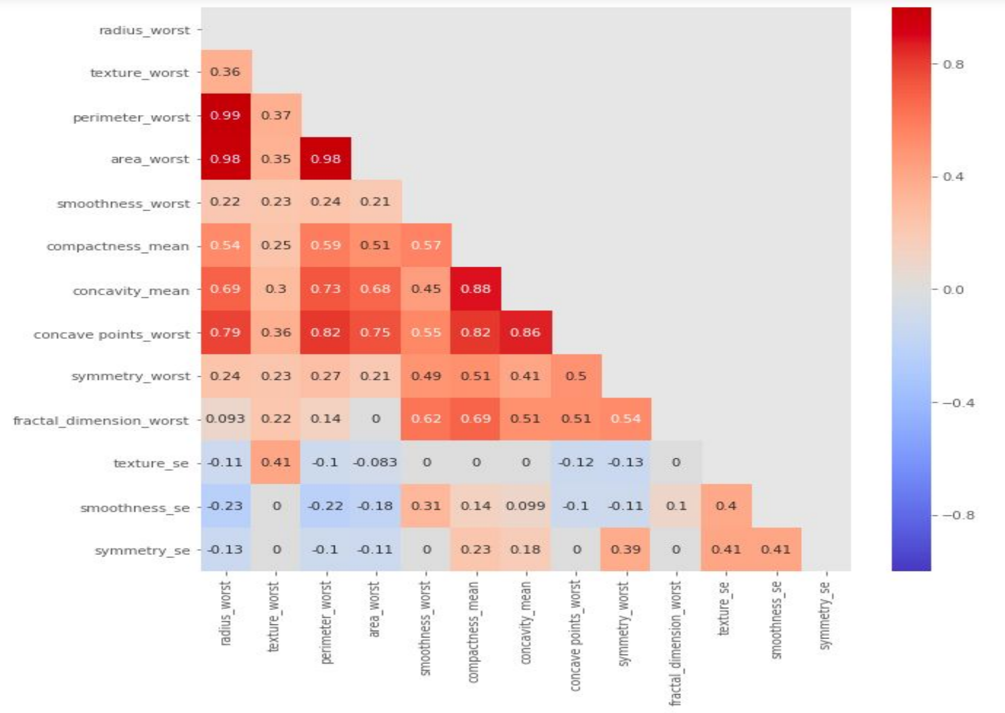
- Used logistic regression to select the best predictor in each group of features between worst, mean, and standard error

- Calculated the correlation between the best feature and the other two features in the same group to confirm that multicollinearity was an issue

- The correlations smaller than .5 between the two features also added to best features list. The final list consists of 13 features

Exploratory Data Analysis

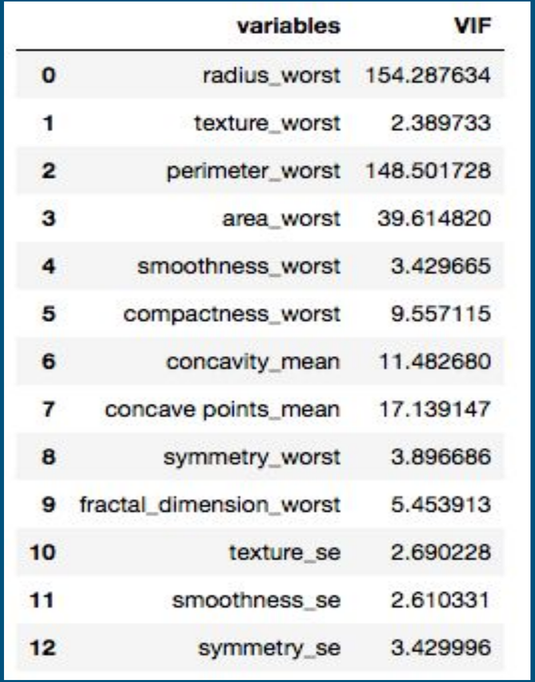
Grayed out non significant correlations between each two features group combination



Investigating Effect of Features on Malignancy

- Radius_worst, perimeter_worst and area_worst have high variation inflation factor because they explain same variance

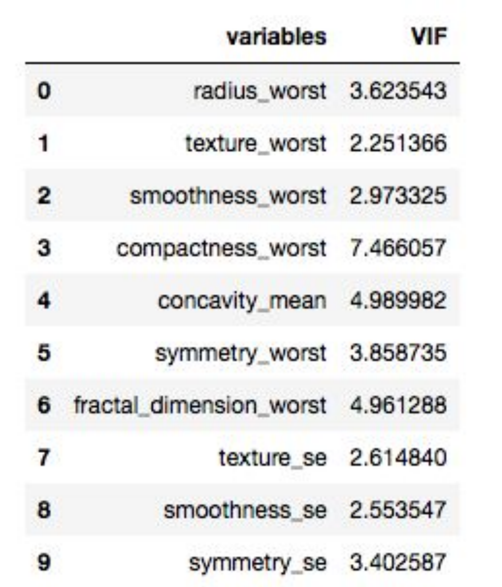
- Also,concavity_mean and concave points_mean are highly correlated and they have high vif



	variables	VIF
0	radius_worst	154.287634
1	texture_worst	2.389733
2	perimeter_worst	148.501728
3	area_worst	39.614820
4	smoothness_worst	3.429665
5	compactness_worst	9.557115
6	concavity_mean	11.482680
7	concave points_mean	17.139147
8	symmetry_worst	3.896686
9	fractal_dimension_worst	5.453913
10	texture_se	2.690228
11	smoothness_se	2.610331
12	symmetry_se	3.429996

Investigating Effect of Features on Malignancy -2

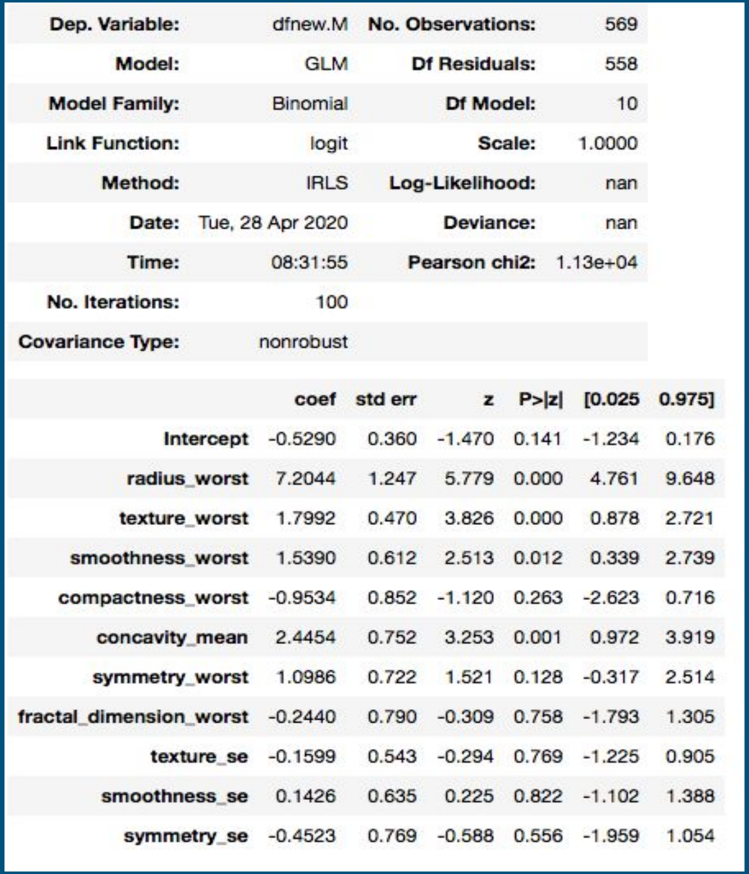
Dropping the features that has the highest variances decreased variation inflation factor notebaly



	variables	VIF
0	radius_worst	3.623543
1	texture_worst	2.251366
2	smoothness_worst	2.973325
3	compactness_worst	7.466057
4	concavity_mean	4.989982
5	symmetry_worst	3.858735
6	fractal_dimension_worst	4.961288
7	texture_se	2.614840
8	smoothness_se	2.553547
9	symmetry_se	3.402587

- Ols report shows that radius is the most important feature with a coefficient more than 3x greater than the next most important feature, concavity, when everything is scaled

- Looking at the p-values, the variables 'radius_worst',’texture_worst”, 'smoothness_worst', and ‘concavity_mean’ ’ seem to be the only significant predictors since their p-values are smaller than 0.05.



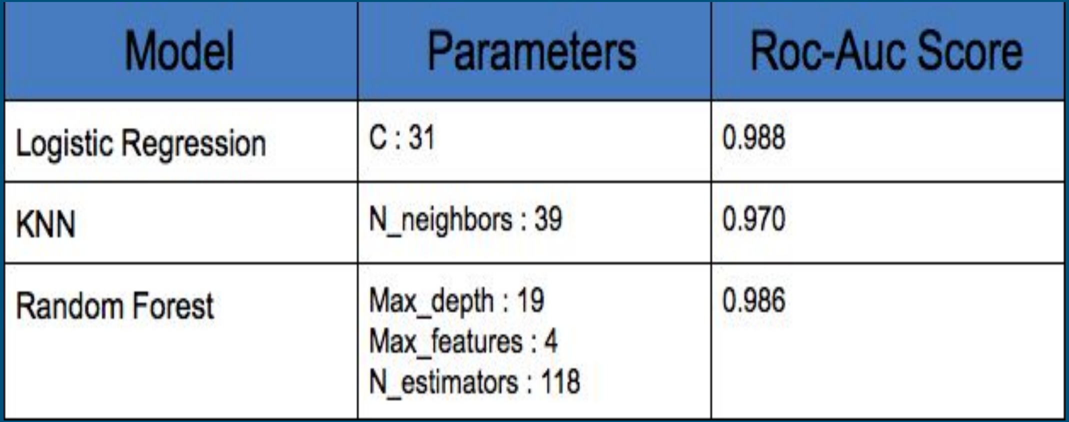
Dep. Variable:	dfnew.M	**No. Observations:**	569
Model:	GLM	**Df Residuals:**	558
Model Family:	Binomial	**Df Model:**	10
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Tue, 28 Apr 2020	**Deviance:**	nan
Time:	08:31:55	**Pearson chi2:**	1.13e+04
No. Iterations:	100		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-0.5290	0.360	-1.470	0.141	-1.234	0.176
radius_worst	7.2044	1.247	5.779	0.000	4.761	9.648
texture_worst	1.7992	0.470	3.826	0.000	0.878	2.721
smoothness_worst	1.5390	0.612	2.513	0.012	0.339	2.739
compactness_worst	-0.9534	0.852	-1.120	0.263	-2.623	0.716
concavity_mean	2.4454	0.752	3.253	0.001	0.972	3.919
symmetry_worst	1.0986	0.722	1.521	0.128	-0.317	2.514
fractal_dimension_worst	-0.2440	0.790	-0.309	0.758	-1.793	1.305
texture_se	-0.1599	0.543	-0.294	0.769	-1.225	0.905
smoothness_se	0.1426	0.635	0.225	0.822	-1.102	1.388
symmetry_se	-0.4523	0.769	-0.588	0.556	-1.959	1.054

Machine Learning

Performed hyperparameter tuning to find out what estimators worked most effectively

Used roc_auc as our metric to determine the best parameters for each model, grid searching over 5-fold cross-validation

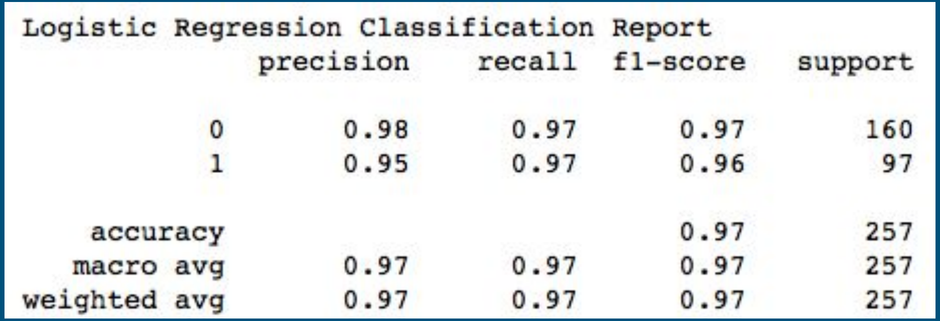


Model	Parameters	Roc-Auc Score
Logistic Regression	C : 31	0.988
KNN	N_neighbors : 39	0.970
Random Forest	Max_depth : 19 Max_features : 4 N_estimators : 118	0.986

Machine Learning

Maximizing recall is more important than precision, although precision is still important

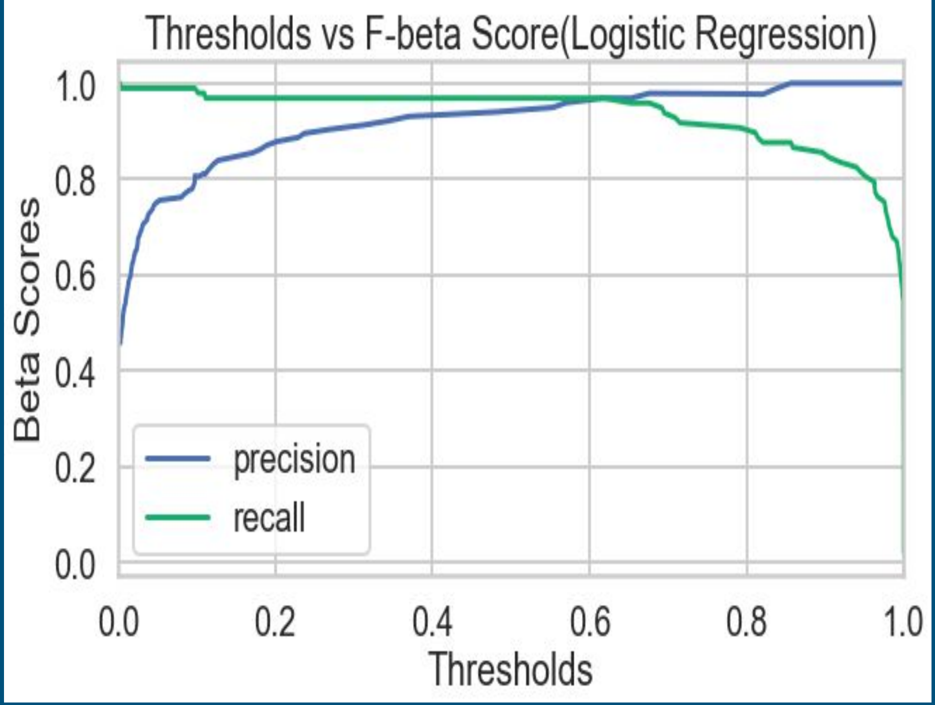
Chose beta as 1,5 to weigh the recall more heavily



```
Logistic Regression Classification Report
              precision    recall  f1-score   support

           0       0.98      0.97      0.97       160
           1       0.95      0.97      0.96        97

    accuracy                           0.97       257
   macro avg       0.97      0.97      0.97       257
weighted avg       0.97      0.97      0.97       257
```



Conclusions

Logistic Regression Classifier with tuned hyperparameters is the optimum algorithm

Feature Selection is important for better score

Able to identify malignant tumors 97% of the time with only 5% false positives

Could try to increase the True Positive rates even higher by using different algorithms or increase the beta