<Predicting breast cancer patients living status after 5 years>

<Yudy Yunardy>

<13 December 2021>

Use Case

- Breast cancer is the most commonly occurring cancer for women and the second most common cancer overall. There were over 2 million new cases in 2018. It is the fifth most common cause of death from cancer in women.

- It is crucial for patients to know the prognosis and estimated lifetime left. It's crucial for doctors to be able to make right prognosis in accordance to patients initial data.

- Today the simple prognosis is usually based on the stage and age and do not consider other factors.

- It is possible to build a system that will make a more accurate prognosis based on all initial patient data with decent accuracy.

Results

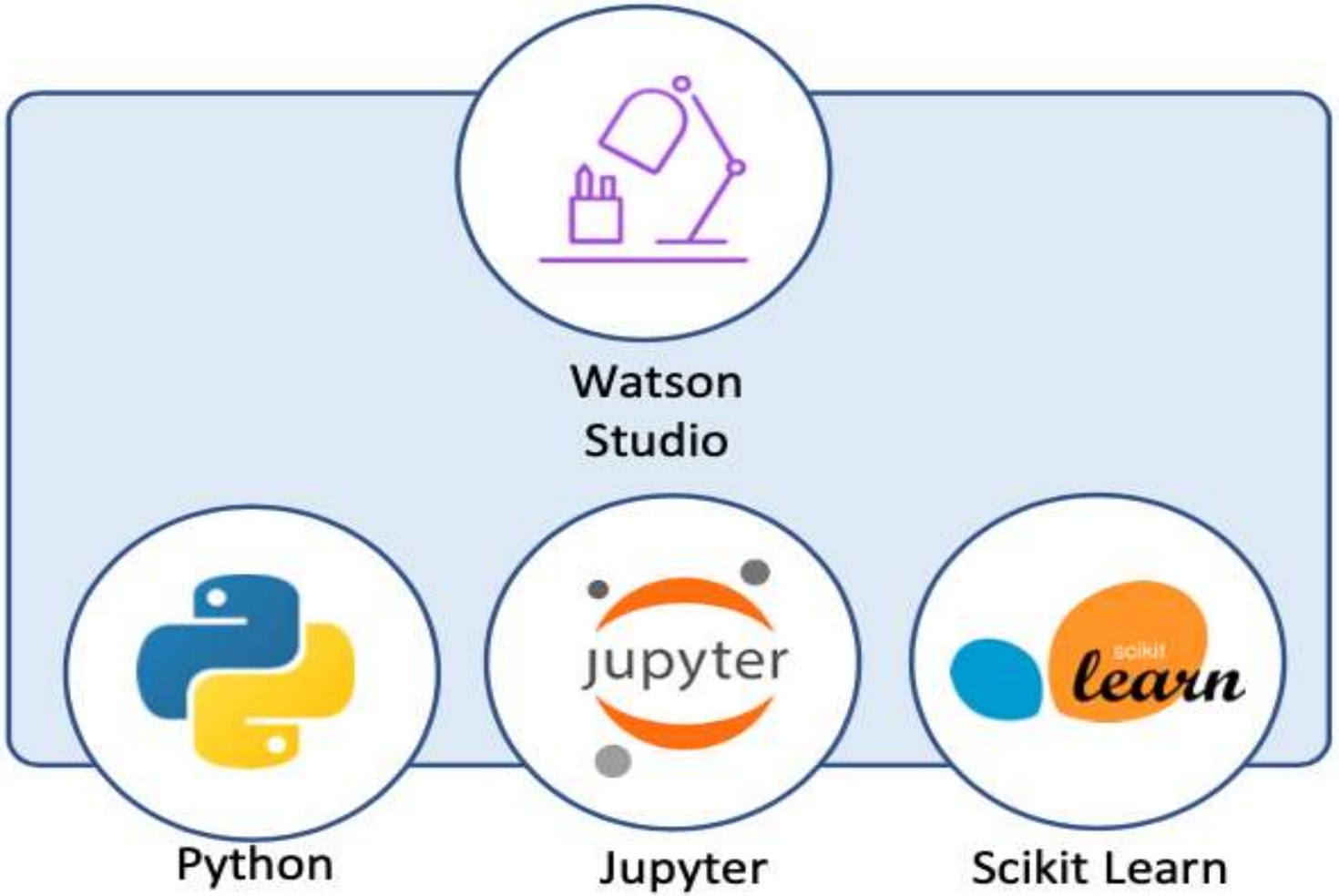
70% accuracy was reached in the patients living status classification after 5 years.

It means that it is possible to make a fully operable system capable of predicting patients living status after a particular amount of years based only on initial medical tests.

It also means that such a system could be used to help doctors in making the right decision about future treatment.

The further work suggested in the direction of enlarging the dataset to achieve higher accuracy and adding additional functionality to the system.

Architecture

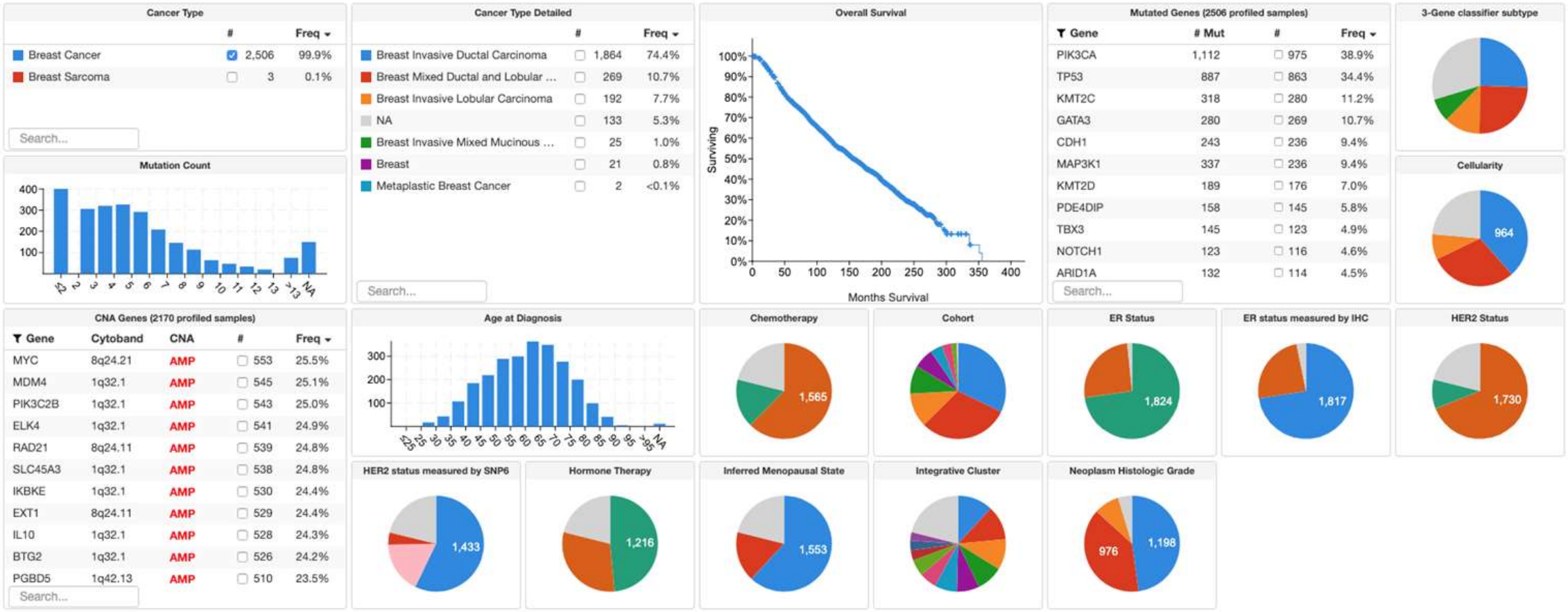


Technologies

- Python
- Jupyter
- Pandas
- Sklearn
- Keras
- Matplotlib, Seaborn
- Apache Spark

Dataset

Breast Cancer Dataset (METABRIC, Nature 2012 & Nat Commun 2016) from The cBioPortal for Cancer Genomics - targeted sequencing of 2509 primary breast tumors



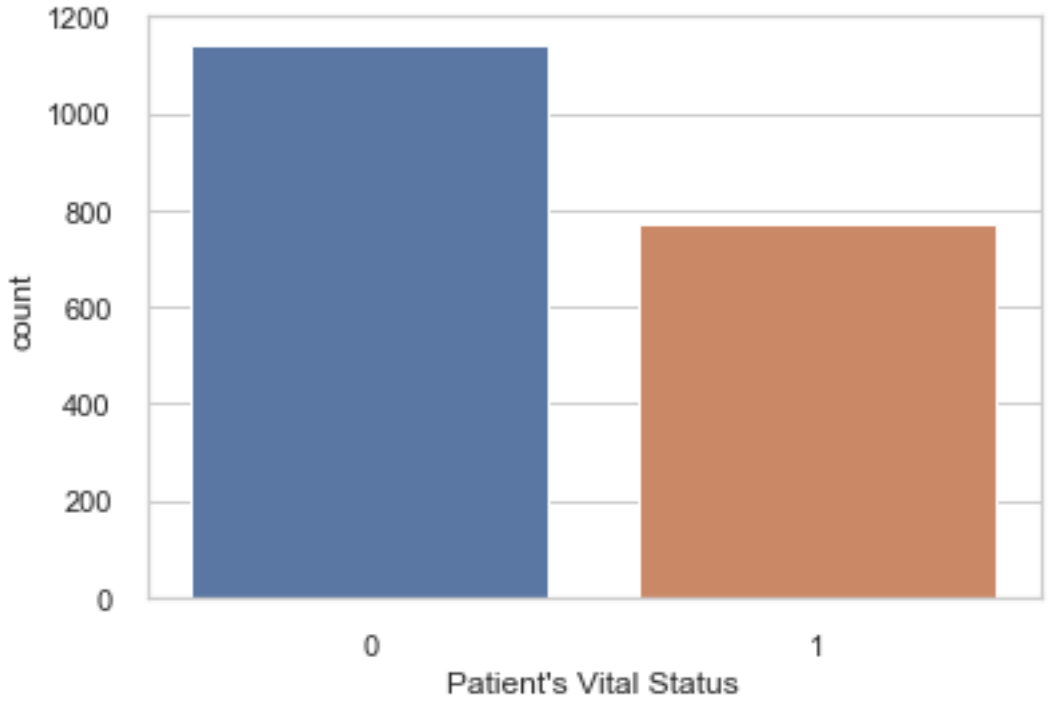
IBM Developer

SKILLS NETWORK

Dataset

Study ID	1980	non-null object
Patient ID	1980	non-null object
Sample ID	1980	non-null object
Age at Diagnosis	1980	non-null float64
Type of Breast Surgery	1954	non-null object
Cancer Type	1980	non-null object
Cancer Type Detailed	1936	non-null object
Cellularity	1916	non-null object
Chemotherapy	1979	non-null object
Pam50 + Claudin-low subtype	1979	non-null object
Cohort	1980	non-null float64
ER status measured by IHC	1937	non-null object
ER Status	1980	non-null object
Neoplasm Histologic Grade	1892	non-null float64
HER2 status measured by SNP6	1979	non-null object
HER2 Status	1979	non-null object
Tumor Other Histologic Subtype	1936	non-null object
Hormone Therapy	1979	non-null object
Inferred Menopausal State	1979	non-null object
Integrative Cluster	1979	non-null object
Primary Tumor Laterality	1869	non-null object
Lymph nodes examined positive	1904	non-null float64
Mutation Count	1859	non-null float64
Nottingham prognostic index	1979	non-null float64
Oncotree Code	1936	non-null object
Overall Survival (Months)	1980	non-null float64
Overall Survival Status	1980	non-null object
PR Status	1979	non-null object
Radio Therapy	1979	non-null object
Number of Samples Per Patient	1980	non-null int64
Sample Type	1980	non-null object
3-Gene classifier subtype	1763	non-null object
Tumor Size	1954	non-null float64
Tumor Stage	1465	non-null float64
Patient's Vital Status	1980	non-null object

- A lot of missing values
- Mixed patients data
- Numerical features are skewed
- Prediction classed are imbalanced



Preprocessing

Data cleaning

Filling missing values

Transformation

Splitting

One-Hot encoding

Scaling

Uniting

Data Cleaning

Study ID	1980	non-null object
Patient ID	1980	non-null object
Sample ID	1980	non-null object
Age at Diagnosis	1980	non-null float64
Type of Breast Surgery	1954	non-null object
Cancer Type	1980	non-null object
Cancer Type Detailed	1936	non-null object
Cellularity	1916	non-null object
Chemotherapy	1979	non-null object
Pam50 + Claudin-low subtype	1979	non-null object
Cohort	1980	non-null float64
ER status measured by IHC	1937	non-null object
ER Status	1980	non-null object
Neoplasm Histologic Grade	1892	non-null float64
HER2 status measured by SNP6	1979	non-null object
HER2 Status	1979	non-null object
Tumor Other Histologic Subtype	1936	non-null object
Hormone Therapy	1979	non-null object
Inferred Menopausal State	1979	non-null object
Integrative Cluster	1979	non-null object
Primary Tumor Laterality	1869	non-null object
Lymph nodes examined positive	1904	non-null float64
Mutation Count	1859	non-null float64
Nottingham prognostic index	1979	non-null float64
Oncotree Code	1936	non-null object
Overall Survival (Months)	1980	non-null float64
Overall Survival Status	1980	non-null object
PR Status	1979	non-null object
Radio Therapy	1979	non-null object
Number of Samples Per Patient	1980	non-null int64
Sample Type	1980	non-null object
3-Gene classifier subtype	1763	non-null object
Tumor Size	1954	non-null float64
Tumor Stage	1465	non-null float64
Patient's Vital Status	1980	non-null object

Age at Diagnosis	1914	non-null float64
Cancer Type Detailed	1875	non-null object
Cellularity	1852	non-null object
Pam50 + Claudin-low subtype	1913	non-null object
ER status measured by IHC	1876	non-null object
ER Status	1914	non-null object
Neoplasm Histologic Grade	1832	non-null float64
HER2 status measured by SNP6	1913	non-null object
HER2 Status	1913	non-null object
Tumor Other Histologic Subtype	1875	non-null object
Inferred Menopausal State	1913	non-null object
Integrative Cluster	1913	non-null object
Primary Tumor Laterality	1809	non-null object
Lymph nodes examined positive	1844	non-null float64
Mutation Count	1802	non-null float64
Overall Survival (Months)	1914	non-null float64
PR Status	1913	non-null object
3-Gene classifier subtype	1702	non-null object
Tumor Size	1892	non-null float64
Tumor Stage	1415	non-null float64
Patient's Vital Status	1914	non-null int64

Filling missing values

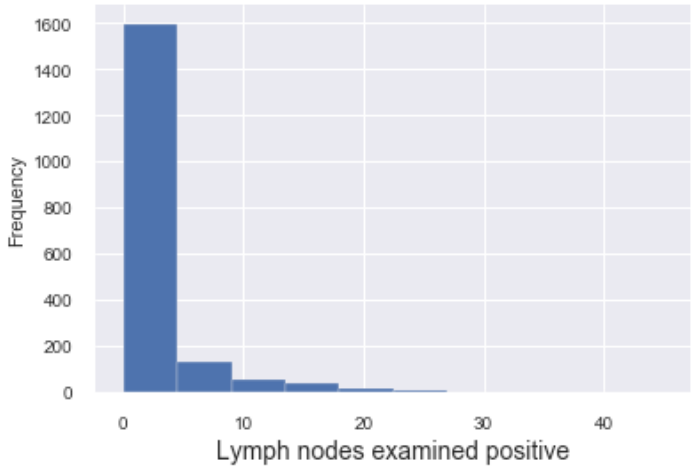
Age at Diagnosis	1914 non-null	float64
Cancer Type Detailed	1875 non-null	object
Cellularity	1852 non-null	object
Pam50 + Claudin-low subtype	1913 non-null	object
ER status measured by IHC	1876 non-null	object
ER Status	1914 non-null	object
Neoplasm Histologic Grade	1832 non-null	float64
HER2 status measured by SNP6	1913 non-null	object
HER2 Status	1913 non-null	object
Tumor Other Histologic Subtype	1875 non-null	object
Inferred Menopausal State	1913 non-null	object
Integrative Cluster	1913 non-null	object
Primary Tumor Laterality	1809 non-null	object
Lymph nodes examined positive	1844 non-null	float64
Mutation Count	1802 non-null	float64
Overall Survival (Months)	1914 non-null	float64
PR Status	1913 non-null	object
3-Gene classifier subtype	1702 non-null	object
Tumor Size	1892 non-null	float64
Tumor Stage	1415 non-null	float64
Patient's Vital Status	1914 non-null	int64

Dropping columns → 1758 rows

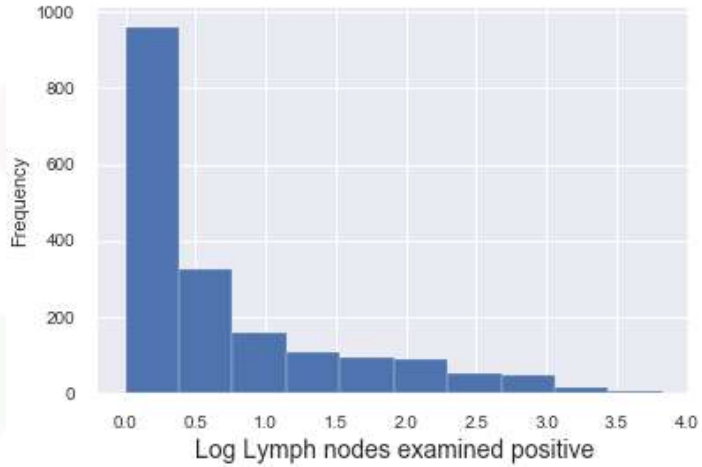
Dropping NaNs → 1120 rows

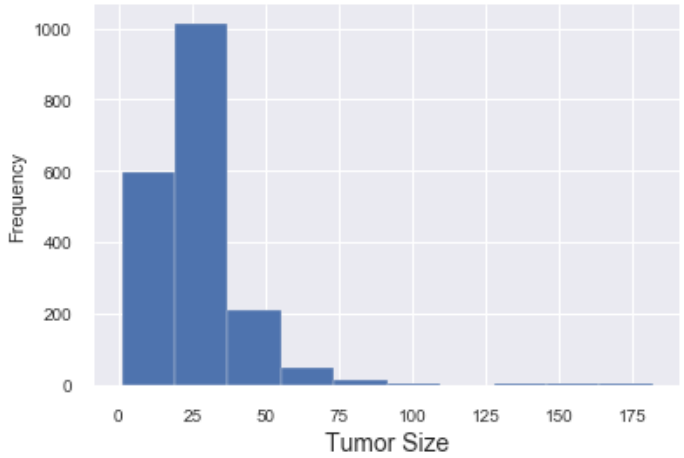
Imputing → 1825 rows

Transformation

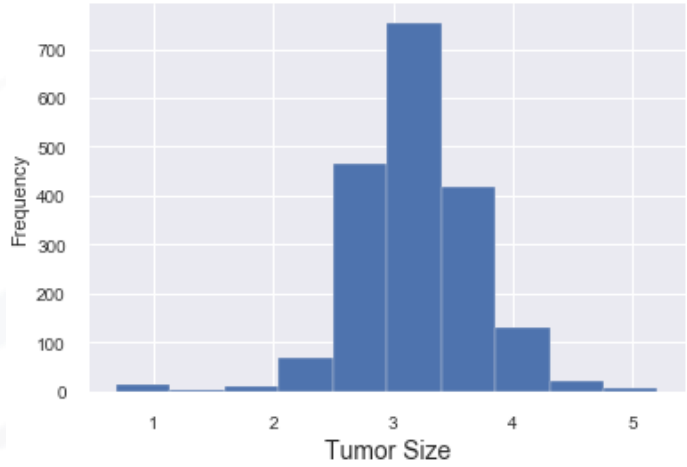


log(x+1)





log(x+1)



Splitting

```
(X_train_num, X_test_num, y_train, y_test) = train_test_split(data_num, labels,
                                                              test_size=0.3,
                                                              random_state=42,
                                                              stratify=labels)

(X_train_cat_oh, X_test_cat_oh) = train_test_split(data_cat_oh,
                                                   test_size=0.3,
                                                   random_state=42,
                                                   stratify=labels)
```

One-Hot Encoding

```
data = df_imputed_transformed
labels = df_imputed_labels
```

```
numeric_cols = ['Age at Diagnosis', 'Lymph nodes examined positive', 'Tumor Size']

categorical_cols = list(set(data.columns.values.tolist()) - set(numeric_cols))
```

```
data_cat = data[categorical_cols]
data_num = data[numeric_cols]
```

```
enc = OneHotEncoder(handle_unknown='ignore', sparse=False)
data_cat_oh = enc.fit_transform(data_cat)
```

Scaling

```
scaler = StandardScaler()

X_train_num_scaled = scaler.fit_transform(X_train_num, y_train)
X_test_num_scaled = scaler.transform(X_test_num)
```

Uniting

```
X_train_scaled = np.hstack((X_train_num_scaled, X_train_cat_oh))
X_test_scaled = np.hstack((X_test_num_scaled, X_test_cat_oh))
```

```
X_train_scaled.shape
```

```
(1277, 59)
```

```
X_test_scaled.shape
```

```
(548, 59)
```

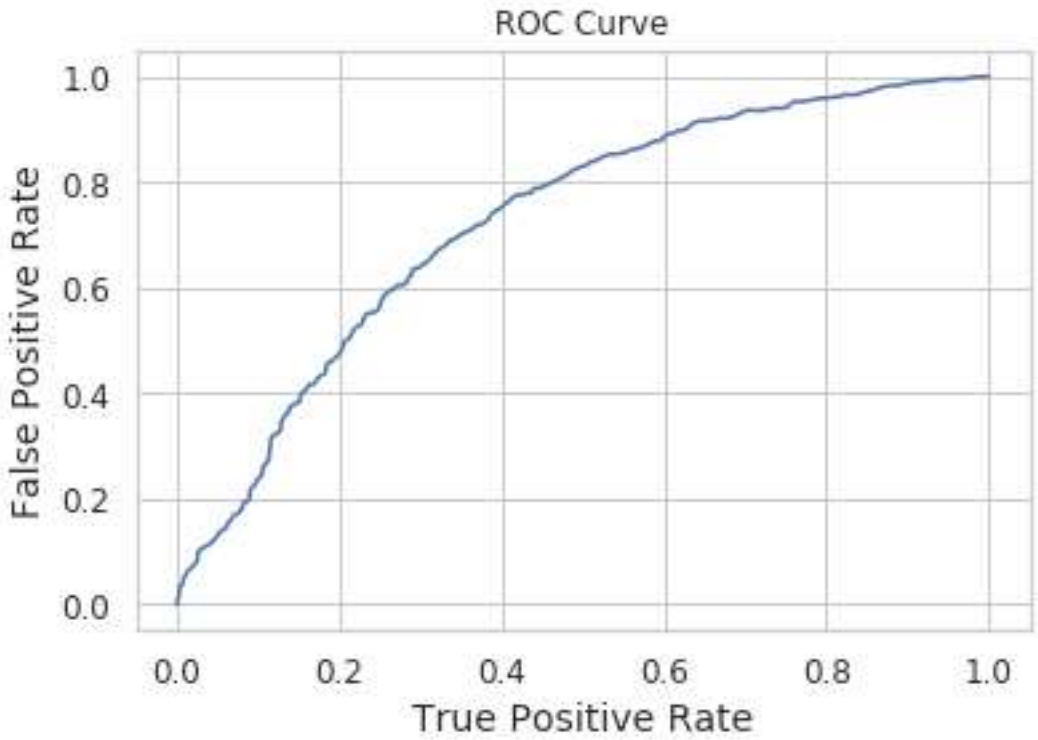
Algorithm selection

Algorithm	Accuracy
Random Forest	64%
Gradient Boosting	69%
Bernoulli Naive Bayes	64%
Logistic Regression	74%
SVM	73%
KNN	65%
MLP	66%
Sequential NN	65%

Iterations

AUROC	Train	Test
Model without normalization	0.705	0.663
Normalization done	0.725	0.686
Feature selection (49 -> 40)	0.724	0.686
Param Tuning	0.775	0.692
Feature Imputation	0.754	0.740

Result




```
roc_auc_score_zeros
```

```
0.7429117024367969
```

```
accuracy_score(y_test, optimizer_zeros.best_estimator_.predict(X_test_scaled))
```

```
0.6916058394160584
```


THANK YOU