Data Science Case Study

Amir Hossein Yazdavar, PhD

NLP Data Scientist Bosch, Sunnyvale
Former Data Scientist:
Weill Cornell Medicine, Cornell University, NYC,
National Institute of Health, Bethesda
{yazdavar}@gmail.com

1 Machine Learning Pipeline

The problem at hand considered as a supervised classification problem with the patient information treated as a training data and target label as class attribute (outcome variable). In this case study, I provided an in-depth analysis of the provided diagnostic measure and their impact on the outcome variables. Then, I fit several common classification algorithm such as Regression (Logistic Regression), Instance-based Algorithms (k-Nearest Neighbor (KNN)), Decision Tree Algorithms (CART), Bayesian Algorithms (Naive Bayes), and Ensemble Algorithms (Gradient Boosting) to fit the predictive model. I defined the machine learning pipeline contains following steps:

- **Data preparation:** 1) Exploratory Data Analysis 2) Data Visualization, 2) Data Selection, 3) Feature Selection, 4) Feature Engineering, 5) Data Transformation
- **Checking the Algorithm:** 1) Perform Measure 2) Evaluate the Performance
- **Improve results:** 1) Algorithm Tuning 2) Ensemble Methods
- Error Analysis: 1) Error correlation 2)Bias and Variance Trade off investigation by drawing Learning Curve
- **Present the model:** Save the model

2 Define Problem: predict the onset of disease based on diagnostic measures

The dataset contains patient's information along with bio-markers that are binary columns with target label. The data scattered through three separated files. The very first step is to concatenate these files while merging all the information with two primary keys "patient id" and "biomarker id". I created a data frame of patients (each row represent a unique patient) contains patient's personal information along with the diagnostic measures (See Figure 1).

	patient_id	institution	cohort_id	gender	age	race	disease_sub_type	comorbidity_index	cohort_qualifier	smoking_status	...	BM15147	BM15148	Bl
0	11156e14a	Saint Penelope Medical Center	14562556998	Male	61.0	White	A	1.0	True	never	...	1	0	
1	113d8066d	Saint Penelope Medical Center	14562556998	Male	62.0	NaN	B	NaN	True	current	...	1	0	
2	113ec3f1	Saint Penelope Medical Center	14562556998	Male	59.0	NaN	A	NaN	True	former	...	1	0	
3	114a37875	Saint Penelope Medical Center	14562556998	Male	68.0	NaN	B	NaN	True	never	...	1	0	
4	11b1d32a1	Saint Penelope Medical Center	14562556998	Male	47.0	White	A	NaN	True	never	...	1	0	

Fig. 1: Create a data frame by integrating patient's information. Each row represent a unique patient.

2.1 Attribute Information:

The data contains demographic information of patient including age(years), gender(categorical), race (categorical), months since diagnosis (numerical), smoking status (categorical), BMI(categorical), exercise frequency (categorical), alcohol usage (categorical), institution information (categorical), biomarker (binary) and target label (Case:1, Control:0) (See Figure 2). The data does not include time information.

```
$ institution           : Factor w/ 6 levels "BioLab, Inc.",..: 5 5 5 5 5 5 5 5 5 5 ...
$ gender                : Factor w/ 2 levels "female","male": 2 2 2 2 2 1 2 2 2 2 ...
$ age                   : num  61 62 59 68 47 65 68 56 29 64 ...
$ race                  : Factor w/ 5 levels "american indian or alaska native",..: 5 NA NA NA 5 3 5 2 3 5 ...
$ disease_sub_type      : Factor w/ 6 levels "A","B","C","D",..: 1 2 1 2 1 1 1 5 1 1 ...
$ smoking_status        : Factor w/ 4 levels "current","former",..: 3 1 2 3 3 1 1 3 3 3 ...
$ months_since_diagnosis: num  16 0 9 0 0 0 33 0 3 2 ...
$ target_label          : Factor w/ 2 levels "Cases","Contorl": 2 2 2 1 2 2 1 1 2 2 ...
$ BM00000               : int  1 0 0 1 0 1 1 1 0 0 ...
$ BM00001               : int  1 1 1 1 0 1 1 1 1 1 ...
$ BM00002               : int  1 1 0 1 1 1 1 1 1 1 ...
$ BM00003               : int  0 0 0 0 0 0 0 0 0 1 ...
```

Fig. 2: Attribute information

3 Exploratory Data Analysis

Exploring the features to get general understanding of the dataset. I am using statistical inference techniques including hypothesis testing with T-test, χ^2 test, Visualization (Probability Distribution Function-PDF), two-way and three-Way Contingency Table(See the Script in R)

3.1 Age Distribution: Visualizing Probability Distribution Function (PDF):

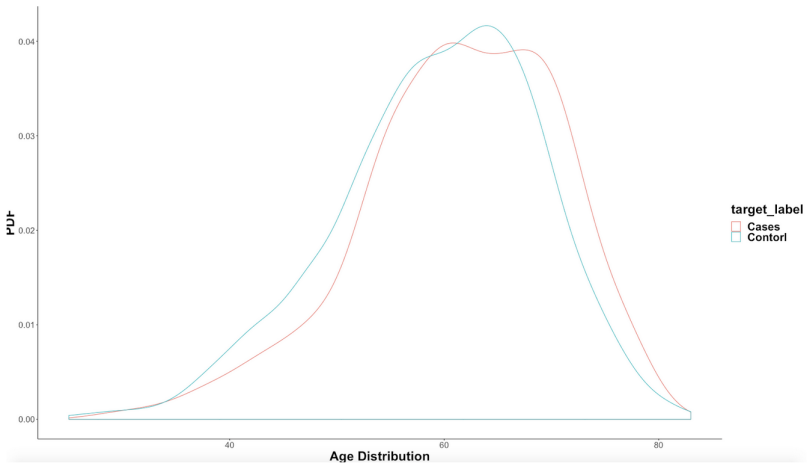


Fig. 3: Age distribution

To better understand how age distribution different for the outcome variable, I study the probability distribution function (PDF) of age with respect to the outcome variable. Figure 3 depicts the age distribution in the dataset for the both case and control class. I performed hypothesis testing with student T-test for exploring whether the difference in the mean of the age of the two groups is significant or not (null hypothesis: age and target variable (case/control) are two independent variable). I observed that there is not a statistically significant difference between the age distribution of control and case (See Figure 4).

3.2 Race, Age, and Gender Distribution

I investigate race, age, and gender distribution to further stratify my understanding of the population of the study. Figure 5 demonstrates that large proportion (47%) of the data coming from white people age between 46-60.

3.3 Month Since Diagnose

Similarly, I study the probability distribution function (PDF) of *Month Since Diagnose* with respect to the outcome variable. Figure 6 demonstrates the month since diagnosis distribution in the dataset for the both case and control class. I performed hypothesis testing with student T-test for exploring whether the

```
> t_test <- t.test(yes_df_sample$age ,no_df_sample$ age )
> p <- t_test$p.value
> p <- p.adjust(p, method = "bonferroni")
> t_test

        Welch Two Sample t-test

data:  yes_df_sample$age and no_df_sample$age
t = 1.8376, df = 708.31, p-value = 0.06654
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0885189  2.6767318
sample estimates:
mean of x mean of y
 61.41499  60.12088

> p
[1] 0.06653673
```

Fig. 4: Statistical significance (t-statistic) of the mean of age features for both classes alpha = 0.05

	(23,34]	(34,46]	(46,60]
american indian or alaska native	0.0000000	0.1584786	0.9508716
asian	0.3169572	2.8526149	10.7765452
black or african american	0.6339144	3.9619651	21.0776545
native hawaiian or other pacific islander	0.0000000	0.0000000	0.9508716
white	0.9508716	10.3011094	47.0681458

Fig. 5: Statistical significance (t-statistic) of the mean of age features for both classes alpha = 0.05

difference in the mean of the month since diagnosis of the two groups is significant or not (null hypothesis: age and target variable (case/control) are two independent variable). I observed that there is a statistically significant difference between the month since diagnosis distribution of control and case. With median of 2 months for control class versus 5 months for case class.

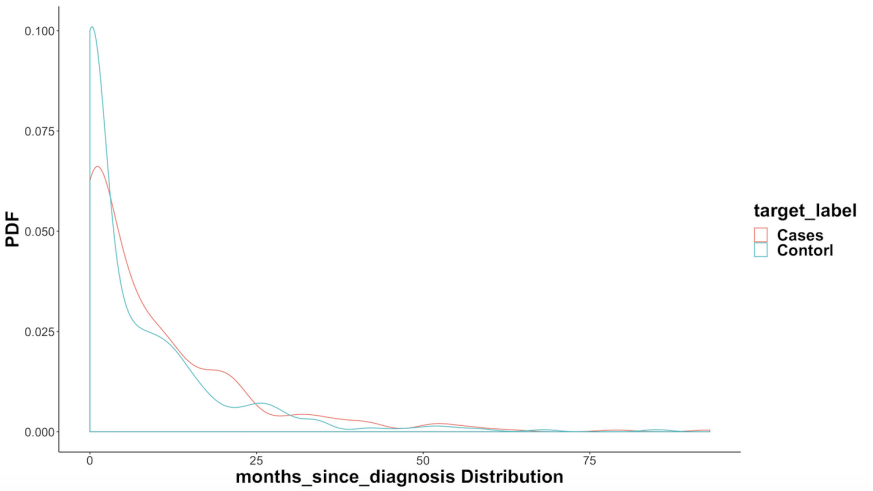


Fig. 6: Month Since Diagnosis Distribution of the both Case and Control Class

3.4 Smoking

Smoking frequency is yet another diagnostic measures of each patient. To investigate the association between smoking frequency and the outcome variable, I performed a chi-square test (null hypothesis: smoking and class label are two independent variables). Figure 8 illustrates smoking association with each of the two classes. Blue circles (positive residuals) show a positive association among corresponding row and column variables, and the red circles (negative residuals) imply a repulsion. My findings indicate a strong association (Chi-square: 25.27, p-value:1.3e-05) between *current* smokers and cases in the dataset and great repulsion for those who never smoked and cases (see red circle in the first row).

Besides, to further investigate the smoking habit among the patients in our dataset. I asked the following question from myself: **How is the distribution of smoking different for the both groups of male and females? how about for different age group?** The dominant group of current smokers are age between 46 to 60 by the 24% whom 14% of them are male and 12% female. Overall male showed more tendency to become a smoker. Smoking is more com-

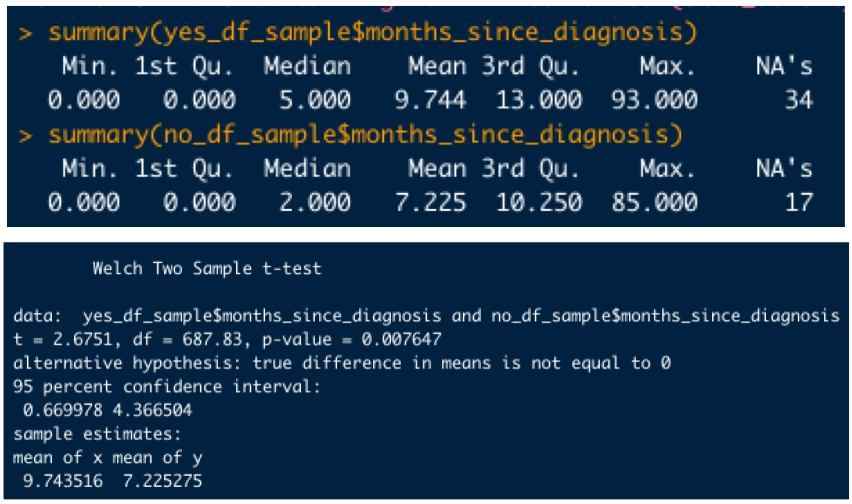


Fig. 7: Month since diagnosis comparison for the both classes. Median and Mean difference.

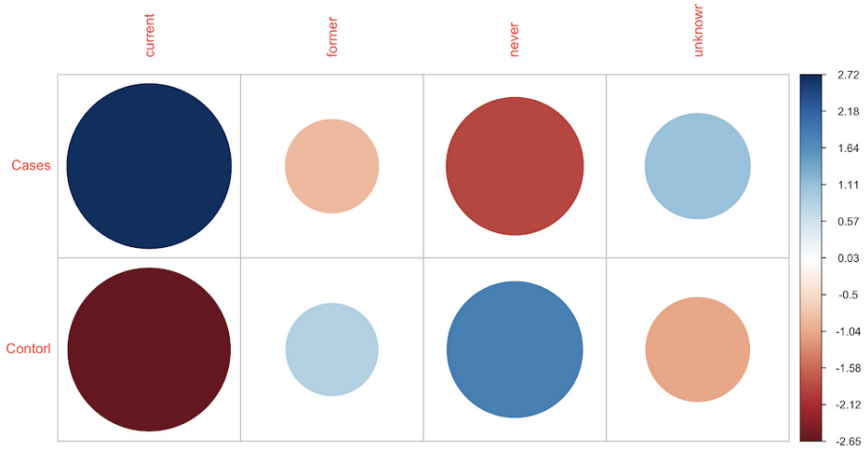
mon among the age group 46-60 irrespective of whether they are classified into case or control group. (See Figure 9).

3.5 Gender

To study the association between gender and target variable, I performed a chi-square test (null hypothesis: Gender and class label (outcome variable) are two independent variables). Figure 10 illustrates gender association with each of the two classes. Blue circles (positive residuals) show a positive association among corresponding row and column variables, and the red circles (negative residuals) imply a repulsion. My findings indicate a strong association (Chi-square: 13.70, p-value:0.00021) between being female and cases in the dataset. Furthermore, 47% of patients are female between 46-60 vs 39% percentage in control class. The control class is more biased towards men.

3.6 Co-morbidity Index

Co-morbidity refers to simultaneous presence of two chronic diseases or conditions in a patient. I explore the impact of this attribute to the outcome variable. It is safe to say that there is a positive correlation between aging and the co-morbidity index (co-occurring of disease) (See Figure 12).




```
> tem_table <- table(merged_df$target_label, droplevels(merged_df$smoking_status))
> tem_table

         current former never unknown
  Cases      125     69   133      20
  Contorl     76     88   188      12
> chisq <- chisq.test(tem_table)
> chisq

        Pearson's Chi-squared test

data:  tem_table
X-squared = 25.276, df = 3, p-value = 1.352e-05

> corrplot(chisq$residuals, is.cor = FALSE)
> p
[1] 0.06653673
```

Fig. 8: Smoking and Class label association (Chi-square test: Color-code: (blue:Association), (red: Repulsion), size: Amount of each cell's contribution).

```
> mytable_yes <- table( df$gender, droplevels(df$smoking_status
> ftable(mytable_yes)
        current former never unknown

female      202    163   310      39
male        229    162   459      47
> prop.table(mytable_yes)*100

           current    former     never   unknown
  female 12.538796 10.117939 19.242706  2.420857
  male   14.214773 10.055866 28.491620  2.917443
```

```
> mytable_yes <- table( df$age_cat, droplevels(df$smoking_status))
> ftable(mytable_yes)
        current former never unknown

(14,19]       0      0     0       0
(19,23]       0      0     0       0
(23,34]       5      2     6       2
(34,46]      31     36    57       9
(46,60]     191    123   303      32
> prop.table(mytable_yes)*100

             current     former      never    unknown
  (14,19]  0.0000000  0.0000000  0.0000000  0.0000000
  (19,23]  0.0000000  0.0000000  0.0000000  0.0000000
  (23,34]  0.6273526  0.2509410  0.7528231  0.2509410
  (34,46]  3.8895859  4.5169385  7.1518193  1.1292346
  (46,60] 23.9648683 15.4328733 38.0175659  4.0150565
```

```
, , target_label = Cases

              age_cat
smoking_status    (14,19]    (19,23]    (23,34]    (34,46]    (46,60]
       current  0.0000000  0.0000000  1.0204082  4.0816327 53.0612245
       former   0.0000000  0.0000000  1.3698630  8.2191781 35.6164384
       never    0.0000000  0.0000000  0.7042254  6.3380282 33.0985915
       unknown  0.0000000  0.0000000  0.0000000  7.1428571 50.0000000

, , target_label = Contorl

              age_cat
smoking_status    (14,19]    (19,23]    (23,34]    (34,46]    (46,60]
       current  0.0000000  0.0000000  2.0408163  4.0816327 35.7142857
       former   0.0000000  0.0000000  0.0000000 12.3287671 42.4657534
       never    0.0000000  0.0000000  0.0000000  8.4507042 51.4084507
       unknown  0.0000000  0.0000000  7.1428571  0.0000000 35.7142857
```

Fig. 9: Smoking distribution between Male and female and different age group.

```
> #--------Chi^2----------
> merged_df <- rbind(yes_df_sample, no_df_sample)
> tem_table <- table(merged_df$target_label, droplevels(merged_df$gender))
> tem_table

         female male
  Cases      207  174
  Contorl    155  226
> chisq <- chisq.test(tem_table)
> chisq

        Pearson's Chi-squared test with Yates' continuity correction

data:  tem_table
X-squared = 13.688, df = 1, p-value = 0.0002159
```

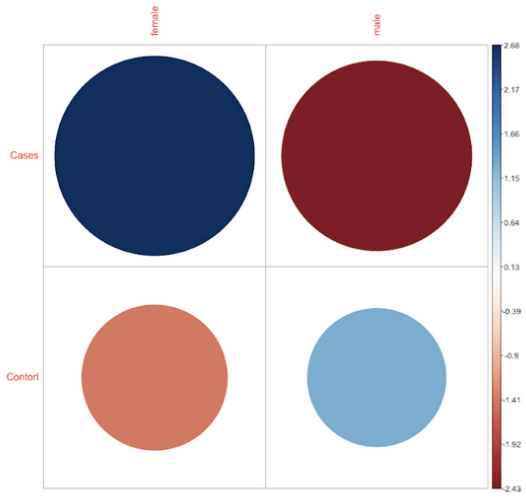


Fig. 10: Gender and Class label association (Chi-square test: Color-code: (blue:Association), (red: Repulsion), size: Amount of each cell's contribution).

```
> mytable_prop
, , target_label = Cases

        age_cat
gender      (14,19]    (19,23]    (23,34]    (34,46]    (46,60]
  female  0.0000000  0.0000000  1.2738854  3.8216561 47.1337580
  male    0.0000000  0.0000000  0.5882353  8.2352941 34.1176471

, , target_label = Contorl

        age_cat
gender      (14,19]    (19,23]    (23,34]    (34,46]    (46,60]
  female  0.0000000  0.0000000  0.0000000  8.2802548 39.4904459
  male    0.0000000  0.0000000  1.7647059  7.0588235 48.2352941
```

Fig. 11: Gender and outcome variable association. Percentage distribution of the outcome variable among different age groups and gender

```
, , target_label = Cases

                age_cat
comorbidity_index   (14,19]   (19,23]   (23,34]   (34,46]   (46,60]
                0  0.000000  0.000000  0.000000  8.571429 42.857143
                1  0.000000  0.000000  0.000000  4.761905 42.857143
                2  0.000000  0.000000  0.000000 11.111111 33.333333

, , target_label = Contorl

                age_cat
comorbidity_index   (14,19]   (19,23]   (23,34]   (34,46]   (46,60]
                0  0.000000  0.000000  0.000000  2.857143 45.714286
                1  0.000000  0.000000  0.000000  9.523810 42.857143
                2  0.000000  0.000000  0.000000 11.111111 44.444444
```

Fig. 12: The association between comorbidity index and aging.

3.7 Exercise/BMI/Alcohol Consumption

I further investigate the other attributes namely Exercise, BMI, and Alcohol Consumption and their relations to the outcome variable. I observe that although there exist some associations between them and the outcome variable but a larger sample size required for drawing a concrete conclusion. The p-value is not small enough in the experiments. For instance, for studying the association between exercise frequency and the outcome variable, I performed a chi-square test (null hypothesis: exercise and class label are two independent variables). Figure 13 illustrates exercise association with each of the two classes. Blue circles (positive residuals) show a positive association among corresponding row and column variables, and the red circles (negative residuals) imply a repulsion. Although the blue circle shows association between control user and those who frequently exercise but the p-value of 0.76 is not big enough to statistically support this hypothesis.

Similarly, for BMI, there exist some associations between BMI and the outcome variable but a larger sample size required for drawing any conclusion and the p-value is not small enough in the experiments (See Figure 14).

For instance, for measuring the association exercise frequency and the outcome variable, I performed a chi-square test (null hypothesis: exercise and class label are two independent variables). Figure 13 illustrates exercise association with each of the two classes. Blue circles (positive residuals) show a positive association among corresponding row and column variables, and the red circles (negative residuals) imply a repulsion. Although the blue circle shows association between control user and those who frequently exercise but the p-value of 0.76 is not big enough to statistically support this hypothesis.

4 Data Preparation

In this section, I am explaining data prepossessing steps I take before fitting the models. I provide clear explanation and justification why each of these steps are necessary.

4.1 Null Variables Exploration

I observed that not all diagnostic measures are available for every patients. Figure 15 shows the distribution of null variables in the dataset. I am removing features that have more than 40% missing in the data as not only they might be misleading for the predictive models, but they are less likely to have any predict power.

4.2 Dropping features with low variance:

Features with low variance do not add much predictive power to the model. First, I removed features that contains only one unique value (variance zero).

```
          HIGH LOW MODERATE NONE
  Cases      5  24       38   11
  Contorl    8  24       32   11
> chisq <- chisq.test(tem_table)
> chisq

        Pearson's Chi-squared test

data:  tem_table
X-squared = 1.1482, df = 3, p-value = 0.7655

> corrplot(chisq$residuals, is.cor = FALSE)
```

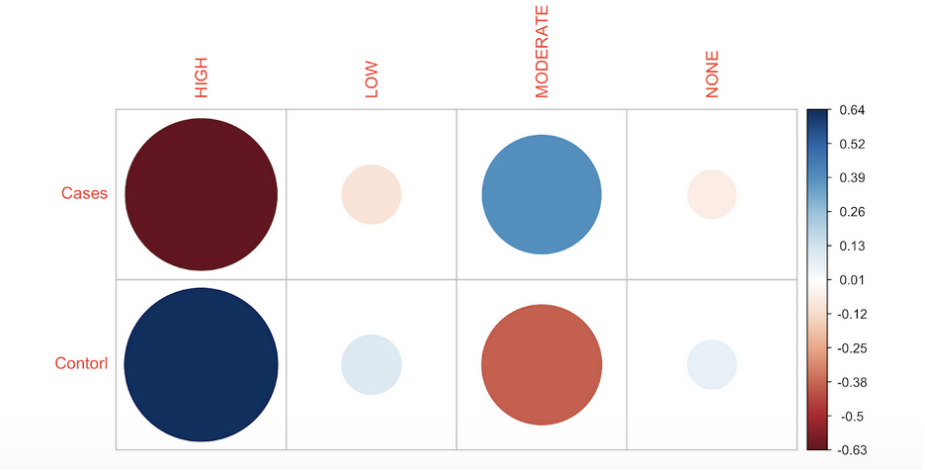
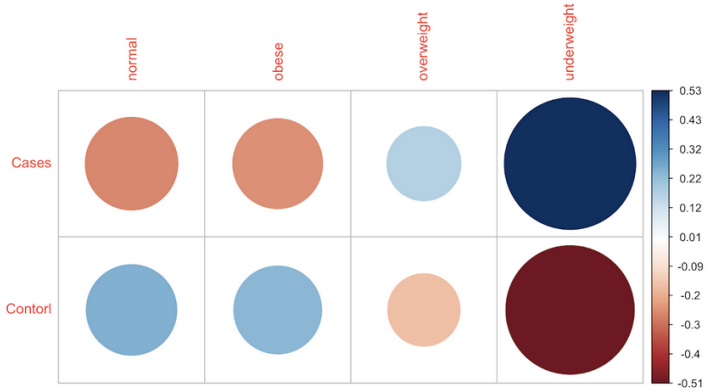


Fig. 13: Exercise frequency and outcome variable association (Chi-square test: Color-code: (blue:Association), (red: Repulsion), size: Amount of each cell's contribution). P-value is not large enough and a larger sample size is required.




```
         normal obese overweight underweight
  Cases      30     2         13           7
  Contorl    36     3         13           5
> chisq <- chisq.test(tem_table)
Warning message:
In chisq.test(tem_table) : Chi-squared approximation may be incorrect
> chisq

        Pearson's Chi-squared test

data:  tem_table
X-squared = 0.85122, df = 3, p-value = 0.8372
```

Fig. 14: BMI and outcome variable association (Chi-square test: Color-code: (blue:Association), (red: Repulsion), size: Amount of each cell's contribution). P-value is not large enough and a larger sample size is required.

```
In [15]:
# check missing values
missing = data.isnull().sum()
missing[missing > 0]

Out[15]:
age                        123
race                       458
comorbidity_index         1381
smoking_status             123
months_since_diagnosis     123
alcohol_usage             1404
exercise_frequency        1404
bmi_level                 1506
BM01069                    696
BM01254                    696
BM01671                    696
BM02181                    696
BM02498                    696
BM03527                    696
BM05569                    696
BM05796                    696
BM05946                    696
BM05948                    696
BM05998                    696
BM06148                    696
BM06675                    696
BM06893                    696
BM07163                    696
BM07181                    696
```

Fig. 15: Null Variable Exploration in the dataset

More than three hundred features removed (324). Removing these features won't harm the predictive model as they are not able to add any predictive power to the final model.

4.3 Imputation

Before performing imputation, I segregate columns contains numerical values from the categorical ones. For numerical features, I am using Median of each column and replace the missing values with them.

For categorical features, I replace the "Nans" values for each column with the Mode of each column. Besides, for categorical features I am using one hot encoding for representing each categories.

4.4 Class Imbalance Problem

Figure 16 illustrates outcome variable distribution of the data. It shows data suffers from class imbalance problem. I am using different re-sampling techniques to address this problem. However, before changing the distribution of data I create a test set contains 30% of data. I am using three common techniques namely up-sampling, down-sampling, and SMOTE for re-sampling the data. Based on the result of preliminary model and the the observation of lack of enough data points, I decided to choose Up sampling as resampling technique.

```
# Class Imbalance problem

         Count Percentage
Cases      381      21.97
Contorl   1353      78.03
```

Fig. 16: Outcome variable distribution shows class imbalance problem.

Create Train/Test Dataset I create a test dataset containing 30% of data. The rest of 70% of data used for fitting the cross validation model. As the data suffers from class imbalance problem, accuracy can be misleading for measuring model's performance. That's why I am using F-1 score for model comparison.

4.5 Dimension Reduction:

Variable selection is the automatic selection of attributes in the data that are most relevant to the predictive modeling problem. Variable selection assists creating more accurate model by choosing features that are most important for the models.

In this dataset, I observe there are a large number of features that may not add any predictive value to the final model. I used variety of Feature selection methods to identify and remove unneeded, irrelevant and redundant attributes which do not contribute to the accuracy of a predictive model. This also helps to avoid over-fitting problem and increase the models' performance.

Dimension Reduction with Variance The figure 17 illustrates the how the variance threshold varies by eliminating input features. It clearly shows lots of features do not add any predictive power to the final model.

```
>Threshold=0.00, Features=14806
>Threshold=0.05, Features=13147
>Threshold=0.10, Features=11366
>Threshold=0.15, Features=9361
>Threshold=0.20, Features=6630
>Threshold=0.25, Features=2
>Threshold=0.30, Features=2
>Threshold=0.35, Features=2
>Threshold=0.40, Features=2
>Threshold=0.45, Features=2
>Threshold=0.50, Features=2
```

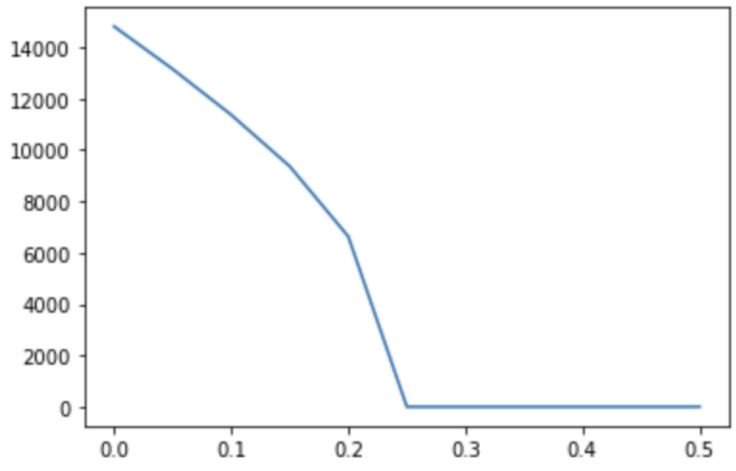


Fig. 17: Plot the threshold vs the number of selected features

Univariate Feature Selection One of the fastest feature selection methods are based on univariate statistical tests. For each feature, measure how strongly the target depends on the feature using a statistical test like χ^2.

I selected 15 salient features namely ['gender male', 'age', 'months since diagnosis', 'BM01029', 'BM04394', 'BM05166', 'BM05719', 'BM05997', 'BM08470', 'BM10516', 'BM11831', 'BM12553', 'BM12577', 'BM14353', 'BM14764'] based on χ^2 experiment. I fit a Random Forrest model on these 15 features and measure the F1 score of 74% on test dataset. The figure 18 shows promising results of the model as a confusion matrix on cross validation and test data set.

```
f1_score is:  0.7336244541484717

Out[61]: <matplotlib.axes._subplots.AxesSubplot at 0x7f8b97de5a00>
```

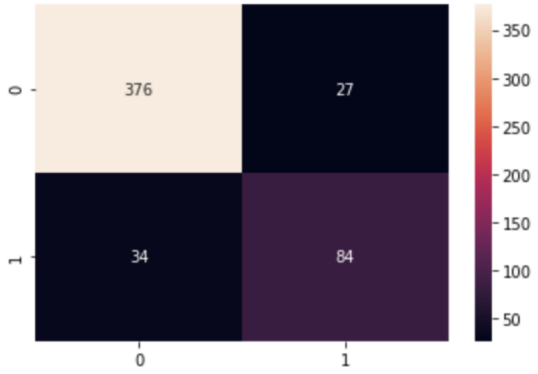


Fig. 18: Confusion Matrix and F1 score of random forest model fitted on 15 salient features based on χ^2 results

Visualizing Principle Components I employed PCA to measure an orthogonal basis in which different individual dimensions of the data are uncorrelated. Before performing PCA, I normalize the data. According to variance ration, 4 components can be chosen (see Figure 19).

Plotting Feature Importance To further extract important features, I used Random Forrest to compare the variable importance among the selected features (See Figure 20).

4.6 Exploring Salient Features Distribution

In this section, I investigate the distribution of the 15 most salient features of the dataset while measuring their skewness, kurtosis, location and variability.

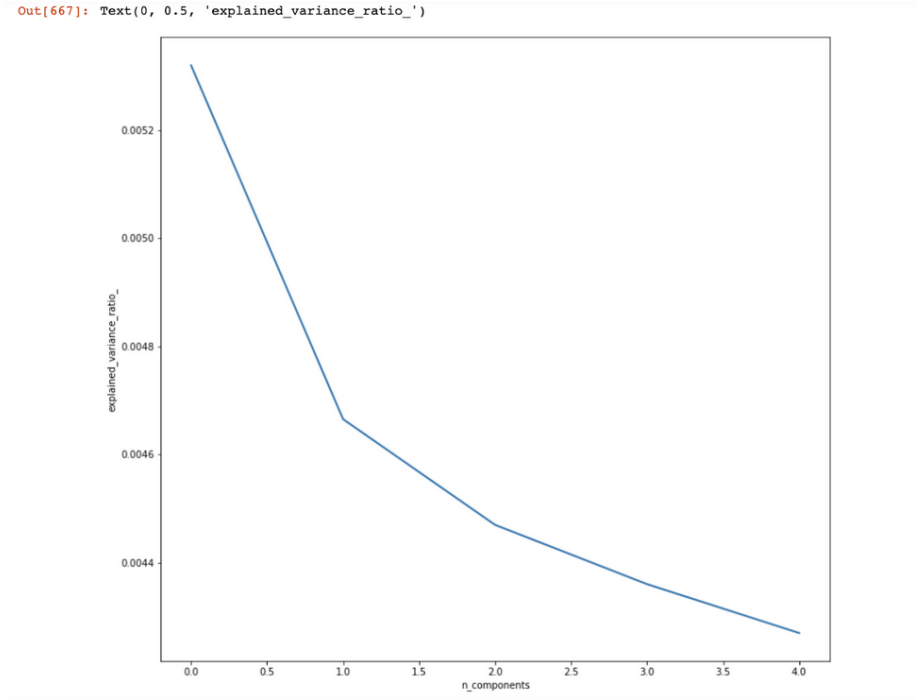


Fig. 19: Uncorrelated dimensions that can explain the data variance

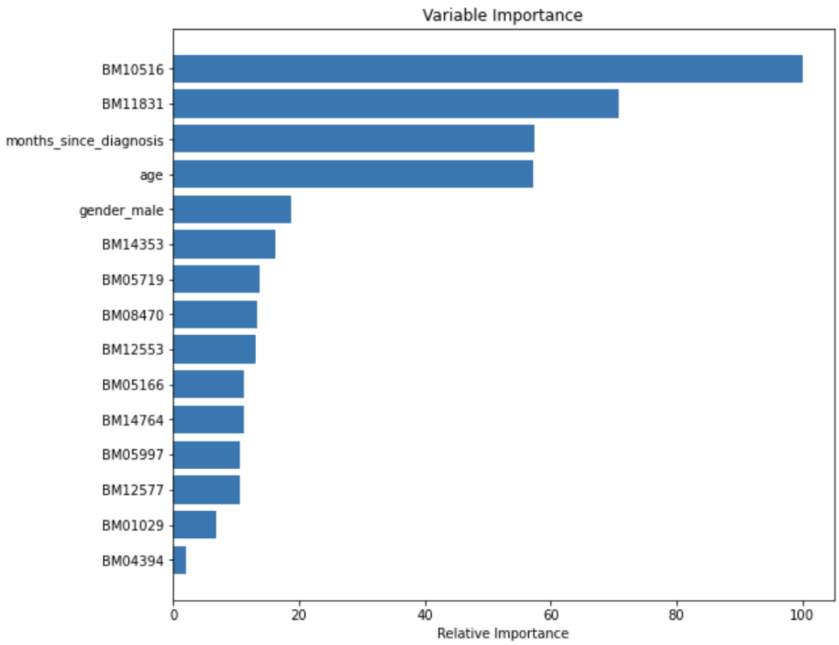


Fig. 20: Ranking features with an ensemble algorithm.

Exploring Location and variability: A fundamental task in many statistical analyses is to characterize the location and variability of a data set. Skewness is a measure of symmetry. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers (see Figure 21, Figure 22, Figure 23, Figure 24).

4.7 Fitting the Models

In this section I explored different machine learning models while comparing their performance based (F1 Score) in 10 Fold Cross Validation and test dataset separately(see Figure 25).

4.8 Data preprocessing

Numerical features preprocessing is different for tree and non tree model. Tree based models does not depend on scaling while Non-tree based models hugely depend on scaling. Some common steps for preprocessing including MinMax Scaler to [0,1] , Standard Scaler (mean = 0 and std =1). I am exploring them and see how our model prediction change (See Figure 26,27,29).

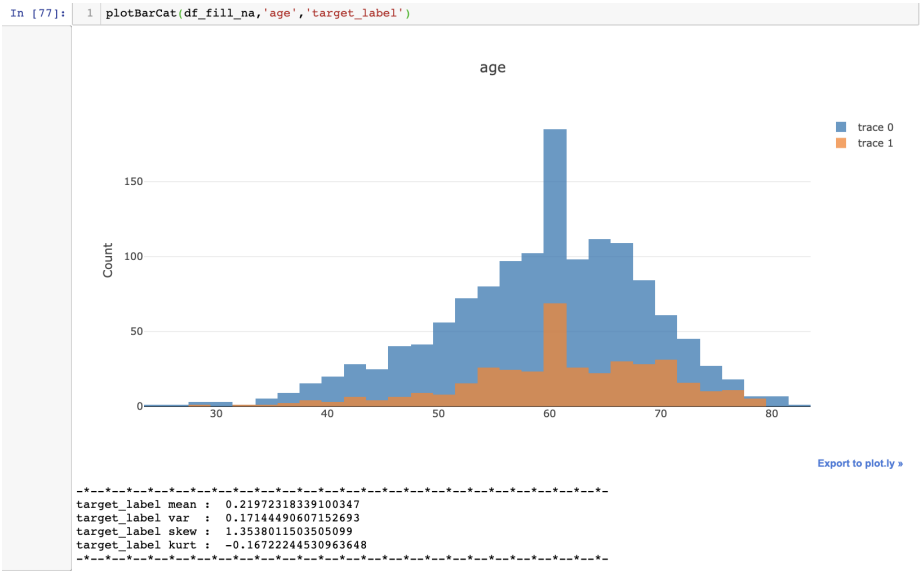


Fig. 21: Exploring Location and Variability of Age Feature with respect to the outcome

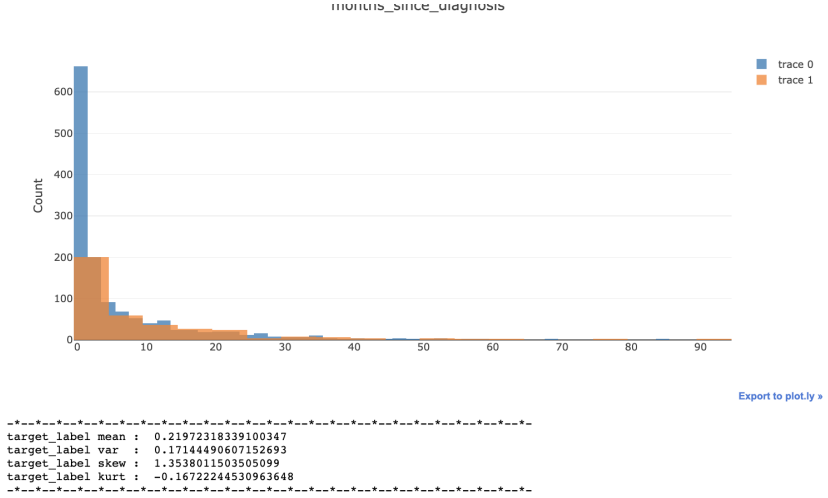


Fig. 22: Exploring Location and Variability of Month Since Diagnosis Feature with respect to the outcome

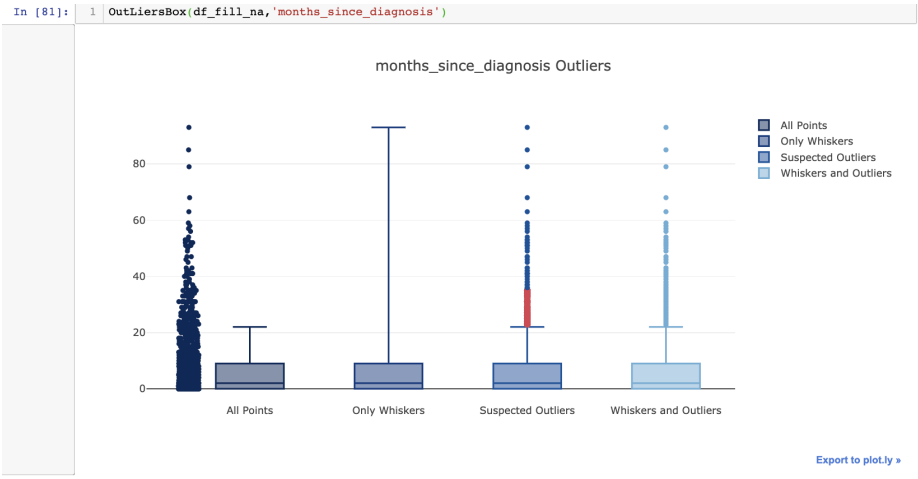


Fig. 23: Outlier Investigation for Month Since Diagnosis Feature

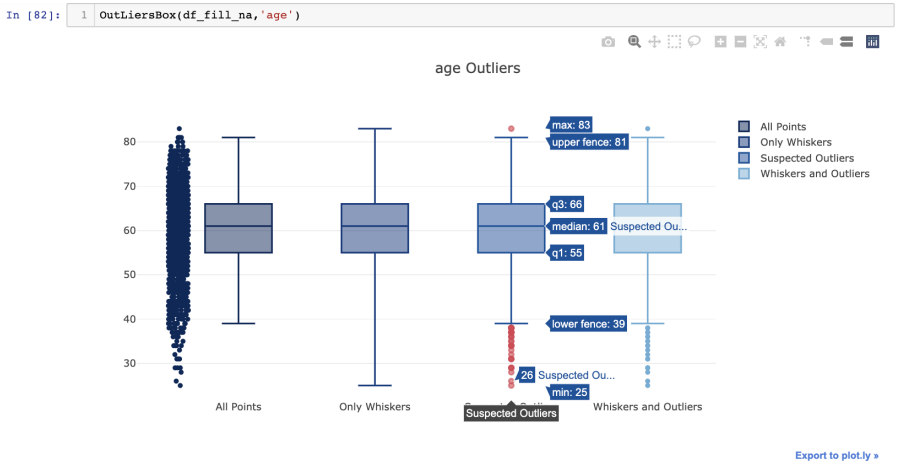


Fig. 24: Outlier Investigation for Age

```
LR: 0.869865 (0.031443)
LDA: 0.859707 (0.030574)
KNN: 0.793757 (0.032447)
CART: 0.943540 (0.013848)
NB: 0.561831 (0.022263)
SVM: 0.620866 (0.034923)
AB: 0.884644 (0.030024)
GBM: 0.912600 (0.014361)
RF: 0.967336 (0.013714)
ET: 0.969979 (0.010016)
```

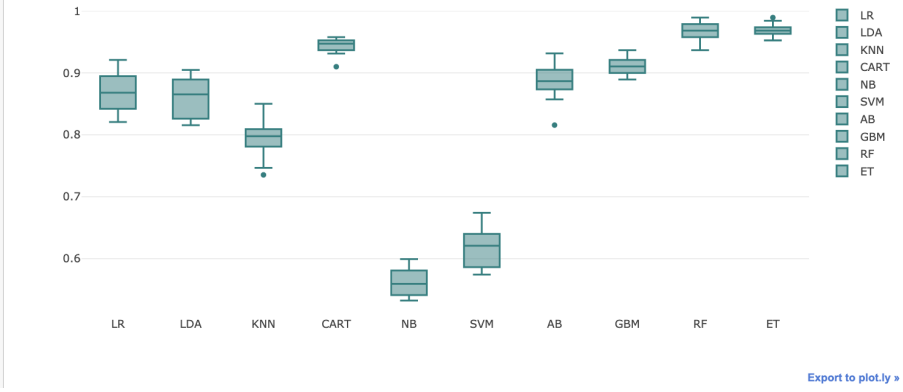


Fig. 25: Model comparison while measuring F1 score in 10 fold cross validation setting

```
standardLR: 0.871986 (0.028084)
standardLDA: 0.859707 (0.030574)
standardKNN: 0.887664 (0.018341)
standardCART: 0.939833 (0.012389)
standardNB: 0.399662 (0.017639)
standardSVM: 0.907849 (0.023104)
standardAB: 0.885170 (0.028834)
standardGBM: 0.912600 (0.014361)
standardRF: 0.968394 (0.011788)
standardET: 0.968394 (0.010811)
```

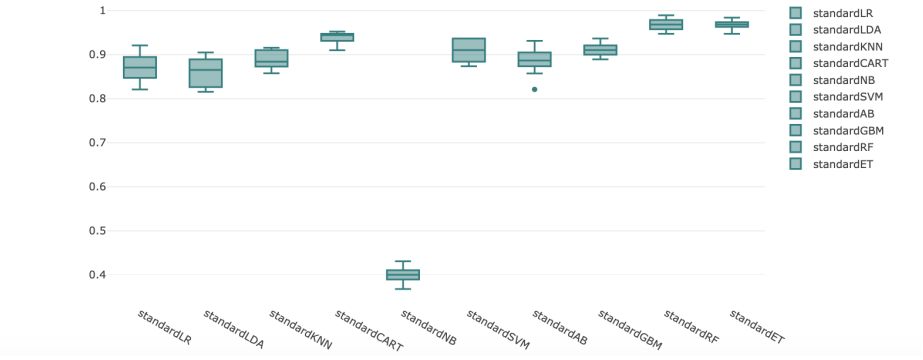


Fig. 26: Model comparison while measuring F1 score in 10 fold cross validation setting with standardizing data

```
minmaxLR: 0.870945 (0.021985)
minmaxLDA: 0.859707 (0.030574)
minmaxKNN: 0.880646 (0.025556)
minmaxCART: 0.947255 (0.013807)
minmaxNB: 0.425637 (0.025683)
minmaxSVM: 0.885724 (0.019964)
minmaxAB: 0.885170 (0.028834)
minmaxGBM: 0.912600 (0.014361)
minmaxRF: 0.968921 (0.012574)
minmaxET: 0.968923 (0.011169)
```

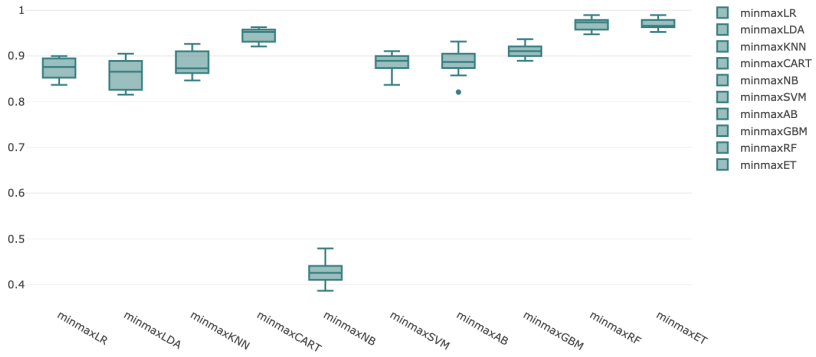


Fig. 27: Model comparison while measuring F1 score in 10 fold cross validation setting after standardizing data with min max preprocessing

```
minmaxLR: 0.863523 (0.023149)
minmaxLDA: 0.860053 (0.026512)
minmaxKNN: 0.851039 (0.032567)
minmaxCART: 0.926694 (0.016390)
minmaxNB: 0.421100 (0.021646)
minmaxSVM: 0.876559 (0.023745)
minmaxAB: 0.879740 (0.027401)
minmaxGBM: 0.917555 (0.022271)
minmaxRF: 0.948325 (0.017384)
minmaxET: 0.953725 (0.017014)
```

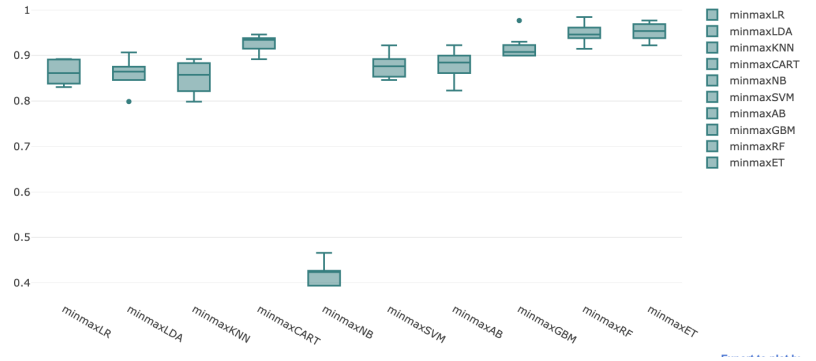


Fig. 28: Model comparison while measuring F1 score in 10 fold cross validation setting after standardizing data with removing outliers from the dataset

Table 29 summarize the impact of data preprcoessing on each model and how F1 score varies in 10 fold cross validation setting after standardizing, Min Max, Outlier removal.

	Model	F_1 Score	Model	F_1 Score	Model	F_1 Score	Model	F_1 Score
0	LR	0.8699	standardLR	0.8720	minmaxLR	0.8709	minmaxLR	0.8635
1	LDA	0.8597	standardLDA	0.8597	minmaxLDA	0.8597	minmaxLDA	0.8601
2	KNN	0.7938	standardKNN	0.8877	minmaxKNN	0.8806	minmaxKNN	0.8510
3	CART	0.9435	standardCART	0.9398	minmaxCART	0.9473	minmaxCART	0.9267
4	NB	0.5618	standardNB	0.3997	minmaxNB	0.4256	minmaxNB	0.4211
5	SVM	0.6209	standardSVM	0.9078	minmaxSVM	0.8857	minmaxSVM	0.8766
6	AB	0.8846	standardAB	0.8852	minmaxAB	0.8852	minmaxAB	0.8797
7	GBM	0.9126	standardGBM	0.9126	minmaxGBM	0.9126	minmaxGBM	0.9176
8	RF	0.9673	standardRF	0.9684	minmaxRF	0.9689	minmaxRF	0.9483
9	ET	0.9700	standardET	0.9684	minmaxET	0.9689	minmaxET	0.9537

Fig. 29: The effect of data preprcoessing on each model and their impact on F1 score in 10 fold cross validation setting after standardizing, Min Max, and Outlier removal

4.9 Parameter Tuning

For each model I performed algorithm tunning while using grid search to come up with the specific parameters that optimize the model. Particularly, for each model we have: **Logistic Regression:** C : Regularization value, the more, the stronger the regularization. Regularization Type: Can be either "L2" or "L1". **KNN**: N neighbors: Number of neighbors to use by default for k neighbors queries. **SVC:** C: The Penalty parameter C of the error term. Kernel: Kernel type could be linear, poly, RBF or Sigmoid. **Decision Tree:** Max depth: Maximum depth of the tree. Row sub sample: Proportion of the observations to consider. Max features: Proportion of columns (features) to consider in each level (double). **AdaBoostClassifier:** Learning rate: Learning rate shrinks the contribution of each classifier by learning rate. N estimators: Number of trees to build.

Figure 30 illustrates the optimized value for each model obtained on the cross validation setting.

4.10 Ensemble Methods

To further enhance the results, I use ensemble methods namely voting and bagging while merging the aforementioned models to gether. Voting is one of the

```
    param = {'C': 3.730229437354635, 'penalty': 'l2'}
    model1 = LogisticRegression(**param)

    param = {'n_neighbors': 1}
    model2 = KNeighborsClassifier(**param)

    param = {'C': 0.7, 'kernel': 'linear'}
    model3 = SVC(**param,probability=True)

    param = {'criterion': 'gini', 'max_depth': None, 'max_features': 7, 'min_samples_leaf': 3}
    model4 = DecisionTreeClassifier(**param)

    param = {'learning_rate': 0.1, 'n_estimators': 200}
    model5 = AdaBoostClassifier(**param)

    param = {'learning_rate': 0.5, 'n_estimators': 250}
    model6 = GradientBoostingClassifier(**param)

    model7 = GaussianNB()

    model8 = RandomForestClassifier()

    model9 = ExtraTreesClassifier()


    models = {'LR':model1, 'KNN':model2, 'SVC':model3,
              'DT':model4, 'ADa':model5, 'GB':model6,
              'NB':model7, 'RF':model8,  'ET':model9
              }
    return models
```

Fig. 30: Parameter Tuning for each model

simplest ways of combining the predictions from multiple machine learning algorithms. It works by first creating two or more standalone models from the training dataset. A voting classifier can then be used to wrap your models and average the predictions of the sub-models when asked to make predictions for new data. I achieved the promising performance of 75% F1 score on unseen dataset.

```
In [200]:
# create the ensemble model
kfold = StratifiedKFold(n_splits=10, random_state=SEED)
ensemble = VotingClassifier(estimators)
results = cross_val_score(ensemble, X_train_sc,y_train_sc, cv=kfold)
print('Accuracy on train: ',results.mean())
ensemble_model = ensemble.fit(X_train_sc,y_train_sc)
pred = ensemble_model.predict(X_test_sc)
print('Accuracy on test:' , (y_test_sc == pred).mean())
f1_scr= f1_score(y_test_sc,pred,pos_label='1')
print('f1_score  on test set is: ',f1_scr)

Accuracy on train:  0.9536842105263158
Accuracy on test: 0.8790786948176583
f1_score  on test set is:  0.7469879518072289
```

Fig. 31: Ensemble model results on cross validation and test dataset

4.11 Error Analysis

Figure 32 illustrates the correlation of error between the models. I observed that errors are significantly correlated, which is to be expected for models that perform well, since it's typically the outliers that are hard to get right.

,

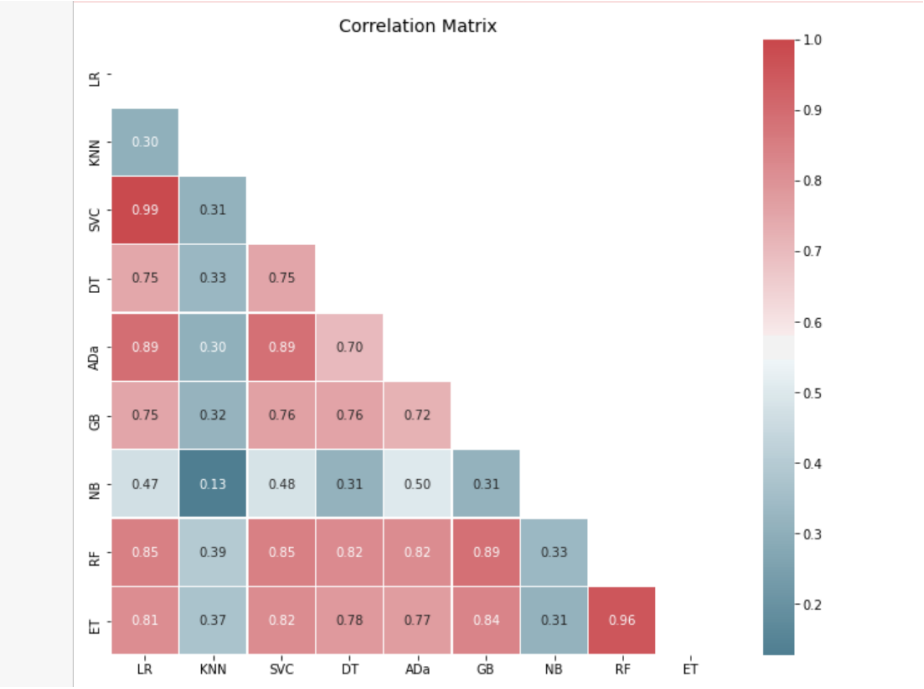


Fig. 32: Ensemble model correlation of error

4.12 Bias and Variance Investigation

To further check bias and variance of the models. I draw learning curve for each model while investigating how bias and variance change after injecting more instance to each model.

The Figure 33 illustrates a learning curve that shows high test variability and a low score up to around 1400 instances, however after this level the model begins to converge on an F1 score. I observe that the training and test scores have not yet converged, so potentially this model would benefit from more training data. Finally, this model suffers primarily from error due to variance (the CV scores for the test data are more variable than for training data) so it is possible that the model is over-fitting.

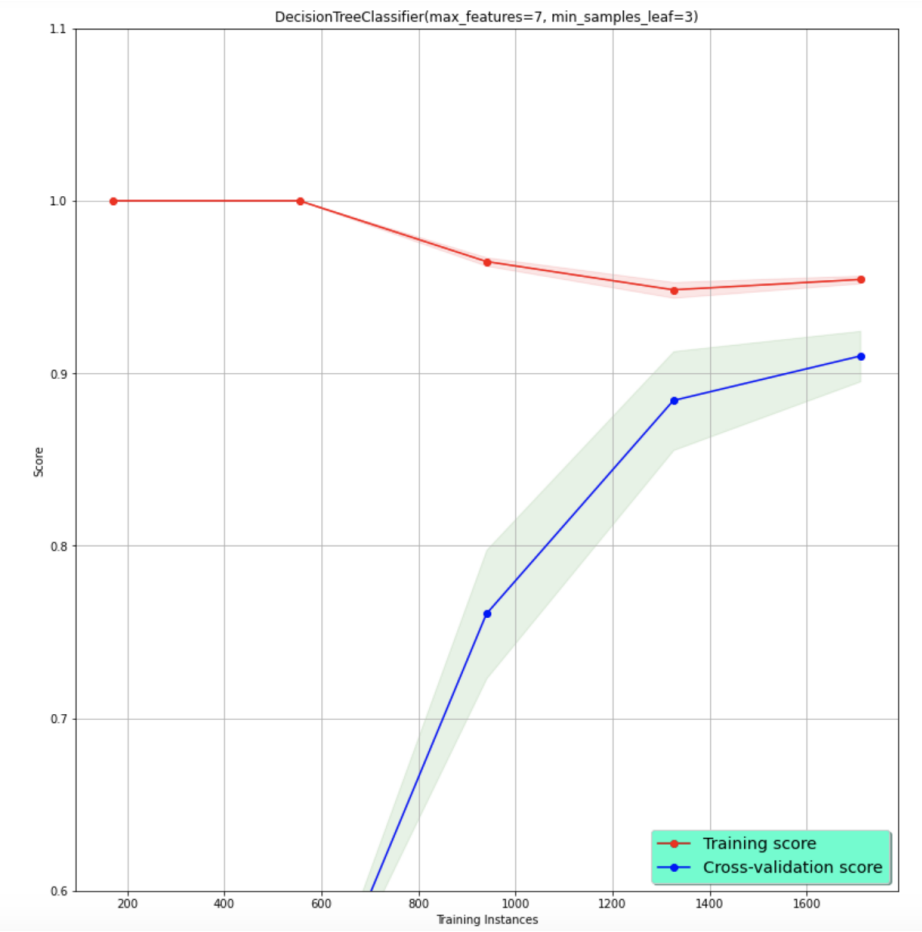


Fig. 33: Decision Tree Bias and Variance Exploration

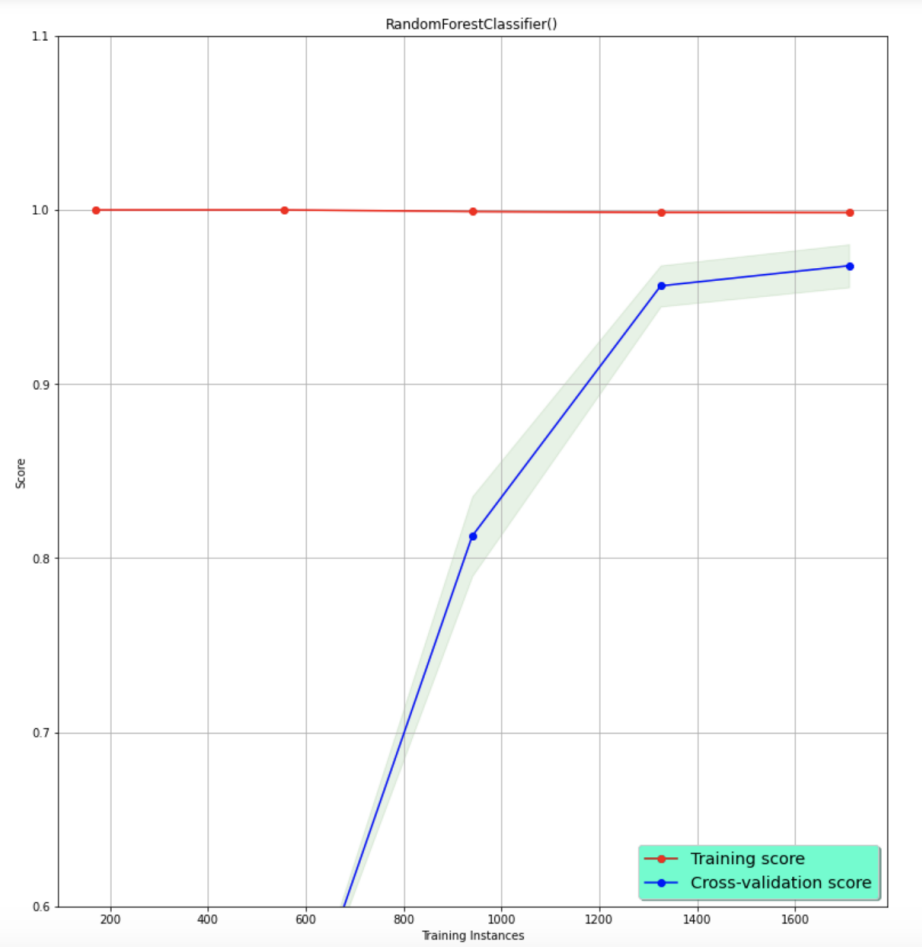


Fig. 34: Random Forest Bias and Variance Exploration

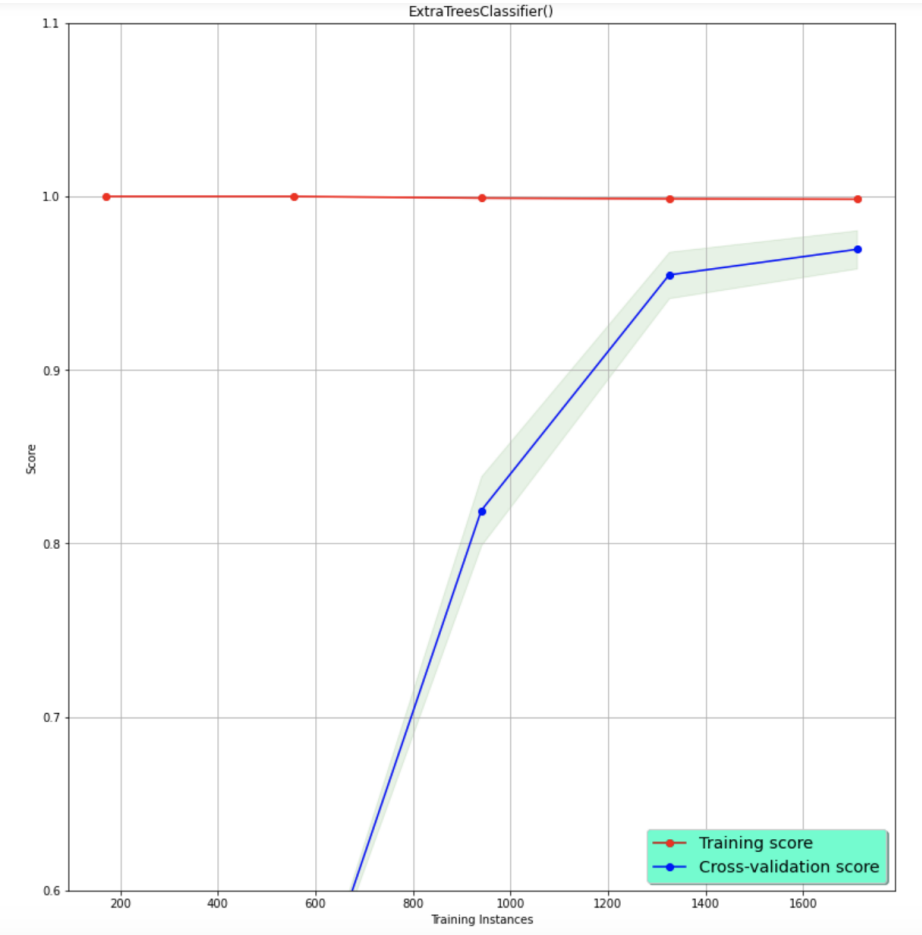


Fig. 35: Extra Tree Bias and Variance Exploration

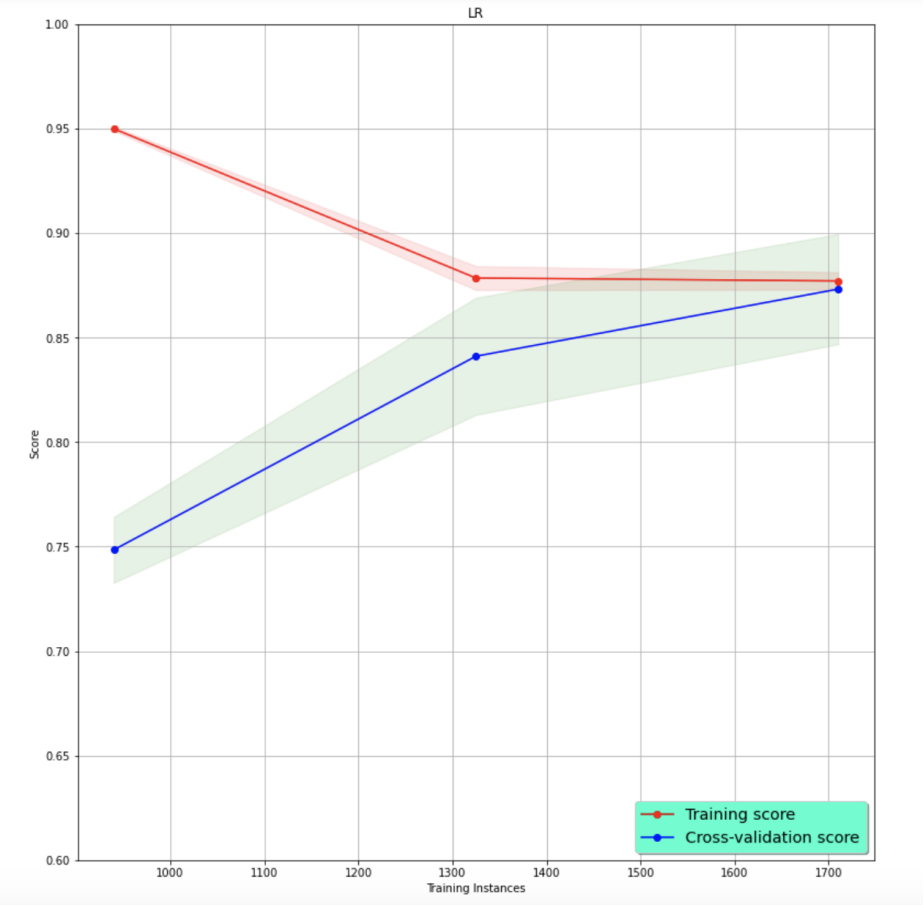


Fig. 36: LR Bias and Variance Exploration

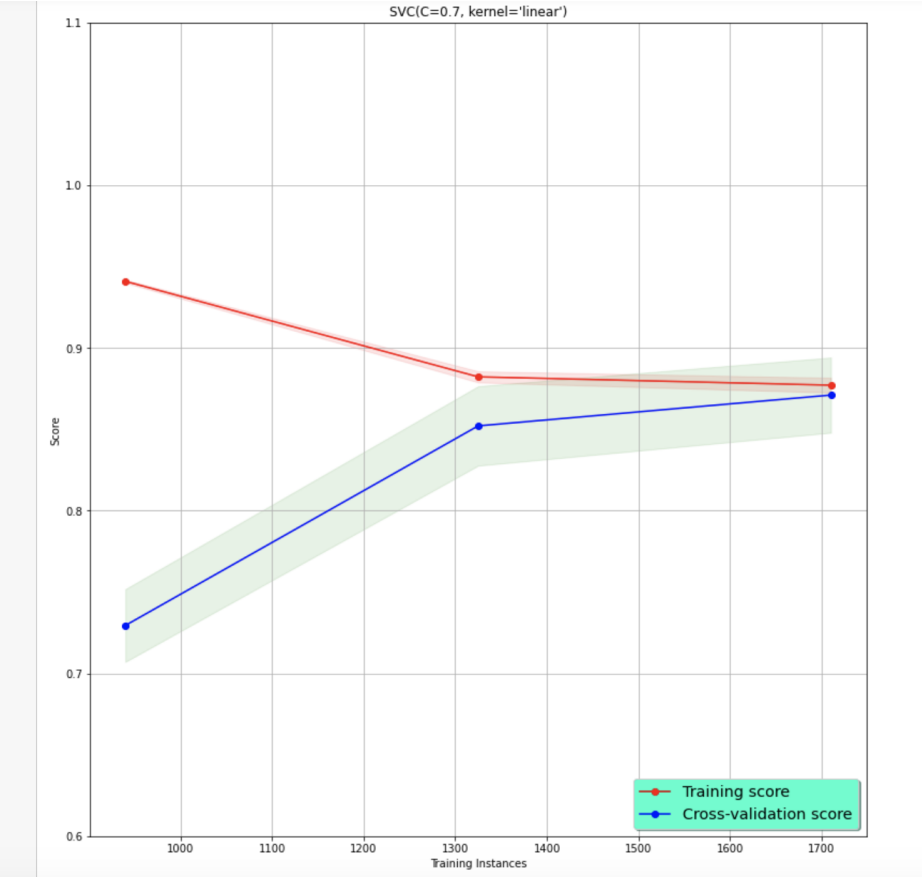


Fig. 37: SVM Bias and Variance Exploration

The learning curve for Extra Tree model (Figure 35) and Random Forrest model (Figure 34) show a very high variability and much lower score until about 1300 instances. It is clear that this model could benefit from more data because it is converging at a very high score. Potentially, with more data and a larger alpha for regularization, this model would become far less variable in the test data.

The learning curve for logistic regression model (Figure 36) and SVM model (Figure 37) show the training and cross-validation scores converge together as more data is added (shown in the left figure), then the model will probably not benefit from more data.