Projecting presence of heart disease from clinical measures with statistical models

Nguyen Hoang The Kiet

Bachelor Student in Data Science and Analytics, College of Humanities and Science, National University of Singapore (Email: e1375598@u.nus.edu).

Keywords: Statistical Modeling, Feature Extraction, k-NN, Decision Tree, Logistic Regression

1. INTRODUCTION

This report proposes a statistical model to predict whether a patient has symptoms of heart disease. The cleaned data was given prior to the analysis, which contains a mixture of categorical and numerical features determining the patient's demographics and clinical measures. Experiments are conducted on three different types of models: k-Nearest Neighbours (k-NN), Decision Trees and Logistic Regression. Models are evaluated based on sensitivity (true positive rate, or TPR) with 5-fold validation, from which the best model from each of its kind is selected and compared based on ROC-AUC *(Receiver Operating Characteristic - Area Under the Curve)* and Precision-Recall tradeoffs.

feature	**type**	**description**	**NA**
sex	`fct(2)`	demographical	0
chest.pain	`fct(4)`	clinical	0
fbs	`fct(2)`	clinical	0
rest.ecg	`fct(3)`	clinical	0
angina	`fct(2)`	clinical	0
blood.disorder	`fct(3)`	clinical	2
age	`int`	demographical	0
bp	`int`	clinical	0
chol	`int`	clinical	0
heart.rate	`int`	clinical	0
vessels	`int`	clinical	0
st.depression	`dbl`	clinical	0
response	**type**	**value**	**count**
disease	`fct(2)`	0 (absence) 1 (presence)	138* 162*

(*) before NA exclusion

Table 1. Dataset features

2. EXPLORATORY DATA ANALYSIS

2.1 Description of dataset

The dataset comprises of 300 raw records and 12 different features. Although the data is already cleaned, there are 2 records with `blood.disorder` $= 0$, noting that no data was recorded in this field. We will simply remove these 2 records from our dataset.

Value	0	1	2	3	Total
Frequency	2	18	163	117	$n = 300$

Table 2. Frequency table of `blood.disorder`

2.2 Categorical variables

There are 6 categorical variables in the dataset, 5 of which are different types of clinical metrics (`angina`, `blood.disorder`, `chest.pain`, `fbs`, `rest.ecg`). In general, there is a disproportionate ratio of categories within the dataset. For example,

- There are *twice* the number of patients who experienced angina induced by exercise (`angina = 0`) compared to those that did not (`angina = 1`);
- The number of patients having high fasting blood sugar level (`fbs = 1`) is about *five times lower* than those having low fasting blood sugar level (`fbs = 0`).

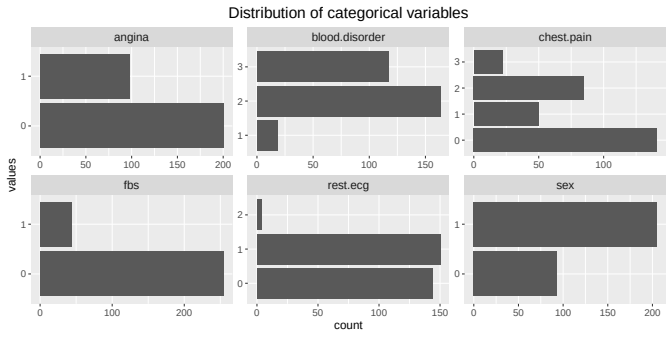

Fig. 1. Distribution of categorical variables

To test the significance of these features in predicting the presence of disease, two methods are used: *percent stacked*

bar chart and *Fisher's exact test*. For a graphical inference, conditional probabilities $\mathbb{P}[Y = y|X_i = x_i]$ (X_i being a categorical feature) are calculated and plotted on a *percent stacked bar chart*. By visual inspection, it can be seen that the level of fasting blood sugar (`fbs`) is weakly correlated to the presence of disease.

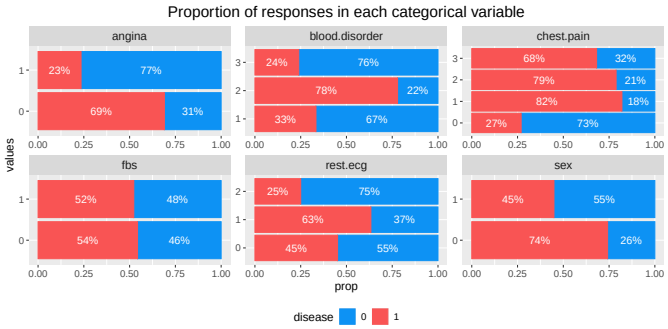

Fig. 2. Relationship of categorical features to the response variable

Indeed, one can calculate the odds ratio as

$$\begin{aligned}\text{OR} &= \frac{\mathbb{P}[Y=1|X=1]}{\mathbb{P}[Y=1|X=0]} \cdot \frac{\mathbb{P}[Y=0|X=0]}{\mathbb{P}[Y=0|X=1]} \\ &\approx \frac{52\% \cdot 46\%}{54\% \cdot 48\%} \approx 0.92\end{aligned}$$

which may suggests an insignificant correlation between fasting blood sugar and the presence of disease.

For a quantitative contigency test, *Fisher's exact test* (Sprent, 2011) is used. From applying Fisher's test, it is observed that there are

- **No evidence of correlation** between the presence of disease and the categorical level of fasting blood sugar ($p > 10^{-2}$).
- **Correlation** between the presence of disease and other clinical measures ($p < 10^{-4}$).

Variable	**Fisher's p-value**
angina	< 0.0001
blood.disorder	< 0.0001
chest.pain	< 0.0001
fbs	0.8703
rest.ecg	0.0019
sex	< 0.0001

Table 3. Fisher's p-value for categorical variables

2.3 Numerical variables

There are 6 numerical variables in the dataset, 5 of which are different types of clinical metrics (`bp`, `chol`, `heart.rate`, `st.depression`, `vessels`). In general, by plotting the histograms for the variables, it is observed that some of the features (`bp`, `heart.rate`, `chol`) are

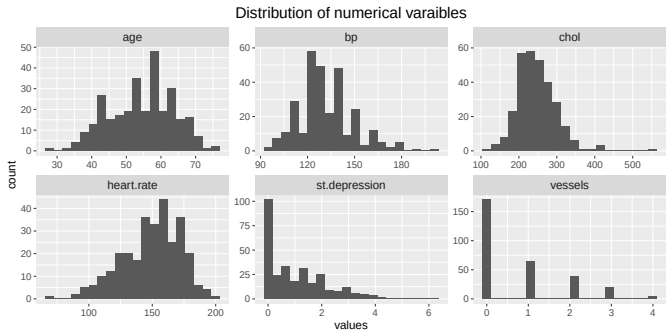

Fig. 3. Histogram of numerical features

'close' to a normal distribution, whilst others are heavily *left-skewed* (`age`) or *right-skewed* (`st.depression`, `vessels`).

The relationship of number variables and the response is tested on a faceted box-plot. By visually comparing the distribution mean of each group, it can be seen that there are

- **Strong correlation** between varibles `age`, `heart.rate`, `vessels`, `st.depression` and the response variable;
- **Weak correlation** between `bp` and `chol` and the response variable

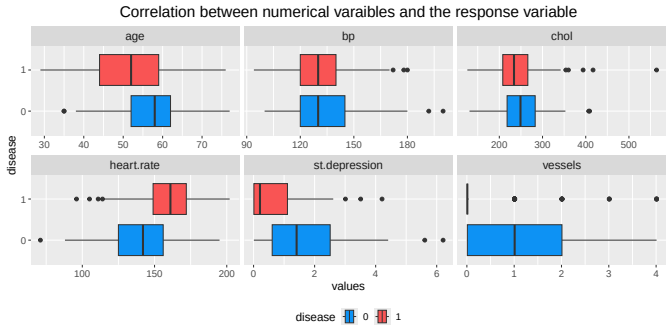

Fig. 4. Boxplot correlation with disease

To test for normality for each variables, a QQ-plot is used for visualisation, along with the p-value from the *Shapiro-Wilk normality test* (Shapiro and Wilk, 1965). As no variables are normally distributed (all $p < 10^{-2}$), normality transformations are also considered. In this case, the transformation of interest is the *Box-Cox transformation*, a parametric transformation that generalises power and logarithmic transformations on positive variables with a single parameter λ:

$$B_\lambda(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln x & \text{if } \lambda = 0 \end{cases}$$

For non-negative variables, a $x \mapsto x + \varepsilon$ transformation, where $\varepsilon = 0.01$, is applied before Box-Cox to prevent the transformation being undefined at $x = 0$.

Different values of λ in the range $[-5, 5]$ are tested, the value of which yields the best *Shapiro-Wilk's p-value* is

retained. It can be observed that `age`, `bp`, `chol` and `heart.rate` are Box-Cox transformable to a normally distributed variable, within a 99% level of significance. In contrast, heavy right-skewed variables (`st.depression` and `vessels`) fail to be normally transformable by the Box-Cox transformation.

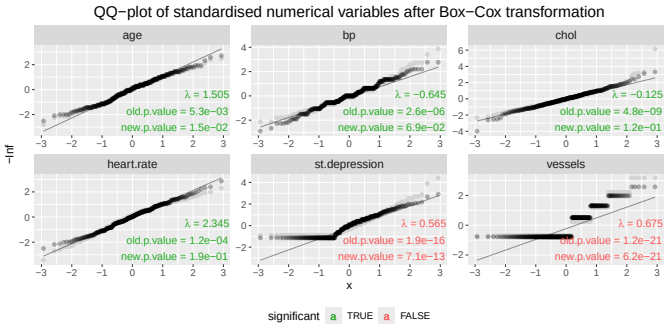

Fig. 5. Best Box-Cox transformation for each numerical variable

2.4 Conclusion

The dataset is a cleaned, 300-record dataset that shows the relationship between demographical and clinical features and the presence of heart disease. A combination of numerical and categorical variables are featured in this dataset, and whilst most features have strong correlation with the response variable, some features seem to be a weak predictor to heart disease (such as `fbs`, `bp` or `chol`). In further sections, models with or without this set of features will be tested against each other.

3. MODEL SELECTION

This report proposes the use of three families of models, *k-Nearest neighbours, Decision trees and Logistic Regression*, in predicting the presence of disease through clinical measures. For each family of model, a range of hyperparameters are validated via 5-fold cross-validation to choose the best candidate. Sensitivity (or True positive rate / TPR) is used as a reference to select the best performing model, with a caution that a high TPR needs not translate to good performance but rather due to underfitting, as the 'True' label would be predicted at a impractical high probability.

To ease the fine-tuning process, a default of threshold of $\delta = 0.5$ is maintained throughout all models.

3.1 k-Nearest neighbours (k-NN)

Methodology. k-Nearest neighbours (k-NN) is a simple, distance-based machine learning classification algorithm that lazily learns its class based on the majority class (for classification) or average value (for regression) of its nearest peers.

Due to k-NN's 'distance-based' nature, categorical features cannot be directly interpreted, hence they need to be transformed to numerical vectors. As in the dataset, most of the variables are numerical or ordinal values, the data is fed into the algorithm *per se*, i.e., by *label encoding*.

Another factor to be concerned of is feature scaling, as different measurements are made in different scales. Two scaling methods are tested: *Normalisation*, where each feature is mapped to the range $[0, 1]$ via a min-max transformation:

$$\mathbf{x} \mapsto \frac{\mathbf{x} - \min \mathbf{x}}{\max \mathbf{x} - \min \mathbf{x}}$$

and *Box-Cox + Standardisation*, where each feature is fed to a Box-Cox transformation with its according optimal parameter λ (if applicable), then standardised:

$$\mathbf{x} \mapsto \mathbf{x}' := \frac{\mathbf{x}^\lambda - 1}{\lambda} \mapsto \frac{\mathbf{x}' - \overline{\mathbf{x}'}}{\operatorname{std} \mathbf{x}'}$$

TPR, 19-NN	**Full**	**Simplified** [1]
Normalisation	86.274%	88.697%
Box-Cox + Standardisation	90.184%	93.188%

[1] `fbs`, `bp` and `chol` are excluded from the model

Table 4. TPR performance of k-NN models, where $k = 19$

Experiments & Results. A total of four models are tested on a various range of values for $k \in [3, 50]$, with a label cutoff at $\delta = 0.5$. It is observed that the simplified model with Box-Cox transformation followed by standardisation performed the best, with TPR = 93.94% saturated at $k \geq 19$.

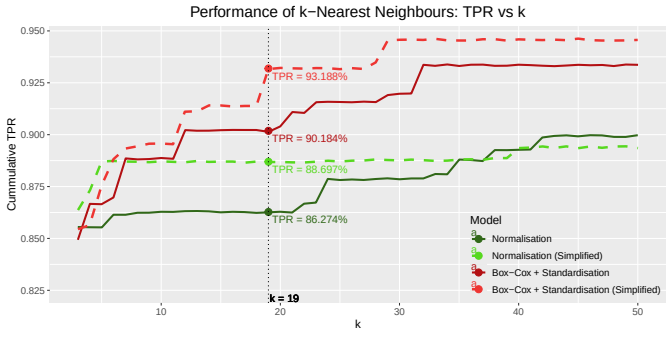

Fig. 6. TPR performance of k-NN models with running k

Generally, standardisation methods outperform normalsation as the best preprocessing method for k-NN. In terms of model complexity, simple models fit best for small values of k, however when k gets larger, then simple models fit the data worse compared to full models; this may rather shows the effect of underfitting than outperformance. By the elbow method, it can be seen that the best performing model is a 19-**nearest neighbour model on a Box-Cox, standardised dataset**.

3.2 Decision tree

Methodology. Decision Tree is a rule-based machine learning algorithm, where its set of rules is generated by recursively splitting the dataset on a chosen feature so that the separation maximises a certain target criterion (e.g. entropy gain, Gini impurity, etc.). Despite being generally less accurate comparing to other machine learning models, it rises in terms of comprehensibility and explainability. As the algorithm does not assume the underlying distribution of the data, transformation such as standardisation or Box-Cox are not required at this stage.

Experiments & Results. Experiments focus in finding the best splitting algorithm, that includes the criterion to be optimised for splitting and the terminating condition. In other words, the following parameters are tested:

- `minsplit`. Number of elements in a leaf node, where the algorithm stops splitting. `minsplit` is let to be $5, 10, 15, \cdots, 100$;
- `split`. Algorithm to determine the optimal split. Either 'Information gain' (Entropy gain) or 'Gini impurity'.

In terms of sensitivity at $\delta = 0.5$, entropy-based algorithms fit well for a low `minsplit`, but as `minsplit` increases, its performance stagnated, both in the full or simplified tree. In contrast, Gini-based trees underperform with low values of `minsplit`, but gradually improve when `minsplit` increases. There is also an observation that simplifing the input does not change the metrics for large `minsplit`s, suggesting the possibility that the algorithm abandons these uncorrelated features at its internal procedures.

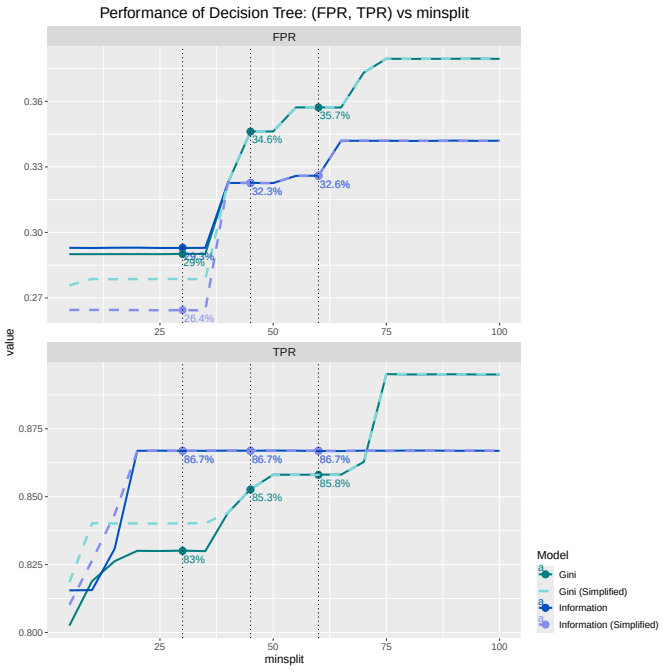

Fig. 7. TPR/FPR performance of decision trees vs. minsplit

One should be cautious, however, when setting a high value of `minsplit`, as it would result in an underfitting tree with most of the datapoint classified as True, optimising TPR at the cost of a high False Positive Rate (FPR). In this example, although a decision tree with `minsplit` $= 60$ produces a 90.0% sensitivity, its speficitivity is a very impractical 43.3%. The tree structure is a simple 2-layer tree that relies only on `blood.disorder` and `chest.pain`, hence the underfitting phenomenon.

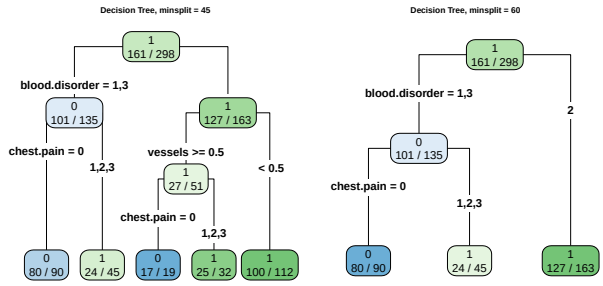

Fig. 8. Left: Moderate 45-minsplit decision tree; Right: Underfitting 60-minsplit decision tree

Choosing the 'best' model amongst all decision trees is rather a nuanced task due to the complex relationship between the `minsplit` hyperparameter with TPR-FPR metrics. For now, these two decision trees are left for further investigation:

- **Information gain,** `minsplit` $= 30$
- **Gini,** `minsplit` $= 45$

3.3 Logistic regression

Methodology Logistic regression is a regression model that estimates the *log-odds* of an event as a linear combination of features:

$$\ln \frac{\hat{p}}{1-\hat{p}} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$$

where $\hat{p}$ is the estimator of $\mathbb{P}[Y = 1]$, $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ are estimates of the linear coefficients. Unlike the usual linear regression, logistic regression has less assumption on its variable. For example, normality of independence variables is not required, hence it is not expected that logistic regression would better-fit on standardised data than the original one.

Experiments & Results Indeed, two logistic models are tested, a full model on all 12 features, and a simplified model on a Box-Cox, standardised version of only 9 significant features. They yields relatively similar results: investigation on the ROC curves produced by the two models shows that they pretty much overlap, except when FPR ≈ 0.125 when the curve of the simplified model concaves down.

It can also be seen from the model coefficients that features contributing the most to the prediction of disease include

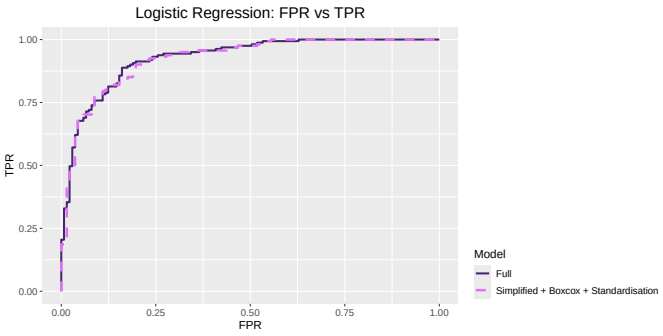

Fig. 9. ROC curve for logistic models

`chest.pain`, `sex`, `blood.disorder` and `vessels`. There is a similarity between logistic models and decision trees, in the sense that these set of features stand as strong predictors for disease.

Feature	Estimate [2]	p-value
chest.pain3	**1.88**	**0.0031**
chest.pain2	**1.86**	**< 0.0001**
sex1	**-1.31**	**0.0085**
blood.disorder3	-1.28	0.0954
chest.pain1	**1.11**	**0.0437**
vessels	**-0.80**	**< 0.0001**
angina1	-0.76	0.0716
st.depression	**-0.69**	**0.0040**
rest.ecg1	0.65	0.0785
heart.rate	**0.53**	**0.0242**
(Intercept)	0.44	0.6332
rest.ecg2	-0.30	0.8992
chol	0.30	0.1504
bp	0.29	0.1213
blood.disorder2	0.21	0.7880
fbs1	0.19	0.7367
age	0.01	0.9661

[#] **bold values** indicate significant features ($\alpha = 0.05$)
[3] sorted by magnitude of coefficients

Table 5. Description of the logistic model

4. MODEL VALIDATION AND COMPARISON

As previous chapter selects the top candidate(s) from three family of models, this chapter describes the full validation of the models and report the most notable results from the process.

The data pipeline starts with a full dataset of $n = 300$ records, two of which with unrecorded fields are then removed. A derivation of the dataset is generated by applying Box-Cox transformation and standardisation to

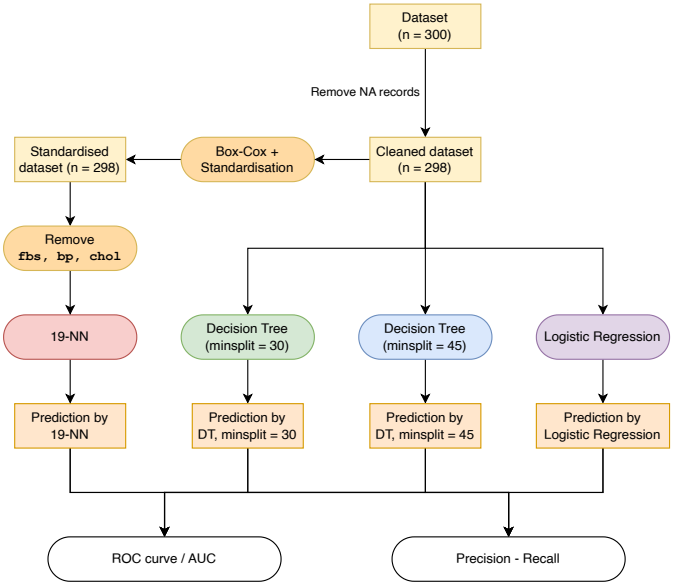

Fig. 10. Process of models validation

all features, regarding categorical values as if they were numerical. The derived dataset is then fed to the 19-NN model for training, whilst other models (logistic, decision trees) are given the cleaned, untransformed dataset. Each model produces a corresponding prediction result, which are then evaluated on the basis of (a) ROC-curve / AUC-score and (b) Precision — Recall curve.

4.1 ROC - AUC (Receiver Operating Characteristic curve and Area Under the Curve)

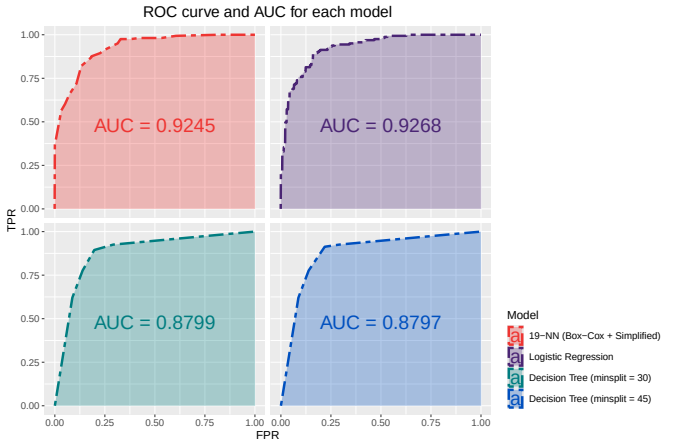

Fig. 11. ROC curves and AUC for each model

Surprisingly, 19-NN and logistic model produce similar ROC curves, despite having different modelling methodologies. It is less so for decision trees, as their alike curves can be explained by the analogous inner mechanisms of their structure.

In terms of AUC score, 19-NN and logistic model give impressive scores of 0.9245 and 0.9268 respectively, whilst decision trees only yield less impressive scores from 0.8797

to 0.8799. The former models excel by considering exhaustive combinations of features, whilst the latter models suffer from overgeneralisation by only considering the most important features and skipping the details.

4.2 Precision - Recall (TPR)

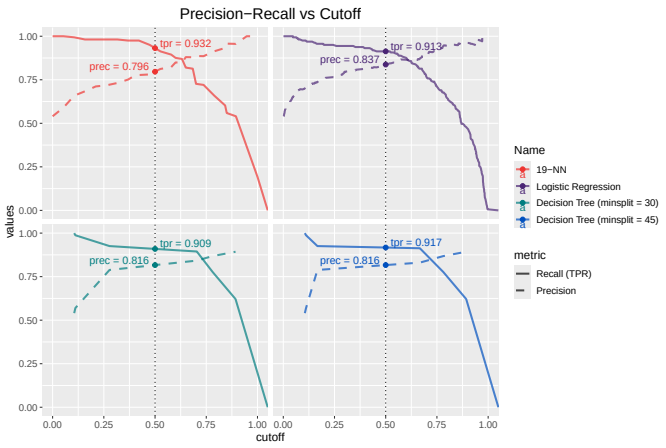

Fig. 12. Percision/Recall – Cutoff curves for each model

As the cutoff threshold increases, precision increases gradually whilst recall (TPR) decreases substantially, especially after the 0.5 threshold. Comparing models at the standard threshold of $\delta = 0.5$, it is observed that no models outperform the others in terms of precision and recall. 19-NN tops at a 0.932 recall but fails at a low 0.796 precision, whilst logistic model observes a humble 0.913 recall but an impressive 0.837 precision. Decision trees are somewhat in between, with the 45-minsplit model having a moderate 0.917 recall and 0.816 precision.

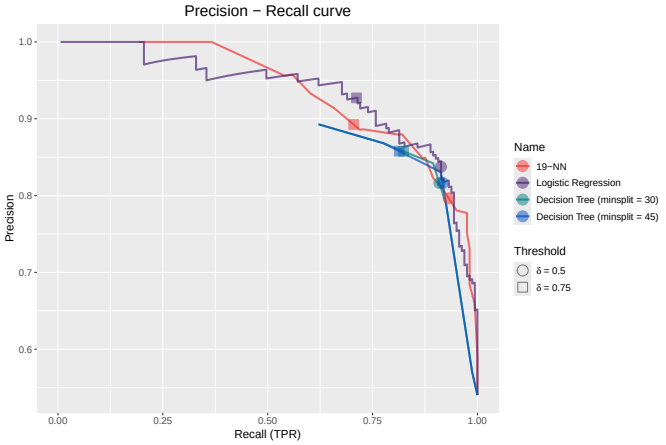

Fig. 13. Percision – Recall curve

As the threshold is moved to $\delta = 0.75$, decision trees perform better in recall whilst logistic model and 19-NN improve in precision. Concaveness around the $\delta = 0.75$ threshold is observed on the precision-recall curve of the 19-NN model, whilst logistic model generally still performs well around this region.

5. CONCLUSION AND COMMENTS

This report illustrates the comparisons between each type of models predicting heart disease based solely on specific demographical and clinical measures. In general, a k-nearest neighbour model, where $k = 19$, and logistic regression model outperform decision trees, with an AUC-score of around 0.93 compared to 0.88. Precision-recall-wise, logistic model shows performance stable lead around the $\delta = 0.75$ threshold, whilst 19-NN lags behind.

For real-life application, **this report would support a logistic regression model**, thanks to its interpretibility and goodness-of-fit compared to other models. Indeed, this model illustrates solid performance in all metrics, with the highest AUC-score and a stable precision-recall curve. Also, by looking at the coefficients of the model, one can deduce the importance of each feature in contributing to the disease status.

This is not present in other experiments. Decision trees, despite its simplicity, fail to give reliable results and a detailed explaination, as factors like `heart.rate` or `angina` are never considered. In contrast, although the best nearest neighbour model perform equivalently to a logistic model, there is no simple explaination on how the model would interpret the features other than a simple distance-based rule. Another drawback of k-NN is *the curse of dimensionality*, as the model is essentially working on a $\mathbb{R}^9$ numerical space in which the difference in distances are negligible. Had the dataset got more features, this effect would have been a significant problem.

For further analysis, this report would suggest that some kind of dimensionality reduction is applied on the data (for instance, Principal Component Analysis, $\mathbb{R}^9 \mapsto \mathbb{R}^2$) so that nearest neighbour models can work on a less-dimensional, compact space. Other models such as Naïve Bayes can be considered for similar small datasets. Finally, one of the big challenges in this report is the size of the dataset, with only $n = 298$ records after cleaning. One could try evaluting the models on a 918-record dataset from the UCI Machine Learning Repository (Fedesoriano, 2021) and would probably produce more reliable results.

REFERENCES

Fedesoriano (2021). Heart failure prediction dataset. URL `kaggle.com/fedesoriano/heart-failure-prediction`.

Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591–611. doi:10.1093/biomet/52.3-4.591.

Sprent, P. (2011). *Fisher Exact Test*. Springer Berlin Heidelberg, Berlin, Heidelberg. doi:10.1007/978-3-642-04898-2_253.