2021

Predicting Emergency Repairs using Classification Method

Tan, J.

Tan, J., Zhang, Q., Sia, W.Y. and Qin, Y. (2021) 'Predicting Emergency Repairs using Classification Method', The Plymouth Student Scientist, 14(2), pp. 465-496.
http://hdl.handle.net/10026.1/18511

The Plymouth Student Scientist
University of Plymouth

*The Plymouth Student Scientist, 2021, **14**, (2), 465-496*

Predicting emergency repairs using classification methods

Jacqueline Tan, Qing Zhang, Sia Wang Ying, Yutong Qin

Project Advisor: Dr Yinghui Wei, School of Computing, Electronics, and Mathematics, University of Plymouth, Drake Circus, Plymouth, PL4 8AA

Abstract

This paper discusses how each explanatory variable affects the possibility of having an emergency repair to people's home with the help of machine learning. Here, the outcome variable is binary. The aim of this is to determine whether increasing the frequency of routine repairs would decrease the frequency of emergency repairs, and the predicted probability of having an emergency repair based on the variable statuses for each property. Data exploratory is first carried out to understand and simplify the dataset obtained from a Housing Association. Statistical models such as logistic regression, decision tree, random forest, linear discriminant analysis and *k*-nearest neighbours are then used to fit the model to the dataset. We also investigate ways to approach the missing values. The best fitted model is then determined by comparing the highest accuracy of the predicted probabilities between these models.

Keywords: Classification methods, emergency repairs, logistic regression, random forest, lda, decision tree, k nearest neighbours, prediction, accuracy

Introduction

The dataset obtained from the Housing Association contains a description of properties and the frequency of each repair to the respective property. The Housing Association is responsible for the provision of approximately 14,200 rented social homes. The data was accumulated as an overall total between the year 2019 and 2020. The explanatory variables are divided into two groups, the categorical variables and the numerical variables (Table 1). Each repairs are categorised into three groups (Table 2). Note that emergency repairs are treated as a binary variable, while planned and routine repairs as numerical variables.

Table 1: Explanatory variables

Categorical variables	Numerical variables
Postcode	Mason (MA)
Date of Construction	Carpenter (CAR)
Bedrooms	Electric (ELE)
Tenants	Electric Door (ELEDOO)
Type	Gas plumber (GASPLU)
SubType	Labourer (LAB)
-	Painter/Decorator (PD)
-	Plasterer (PLA)
-	Plumber (PLU)
-	Window repair (WINREP)

Table 2: The type of groups comprising of different time frames for repairs to be completed.

Group	Description
Emergency repairs (24H)	To be completed within 24 hours
Planned repairs (60D)	To be completed within 60 working days
Routine repairs (3D, 20D)	To be completed within 3 days and 20 working days

The aim of the investigation is to determine whether increasing the frequency of routine and planned repairs to people's home would decrease the possibility of having an emergency repair, which represents as the outcome (dichotomous) variable.

This is a case of a classification problem as we are predicting a qualitative response for an observation. It involves assigning the observation to a category, or class. Statistical models such as logistic regression, decision tree, random forest, linear discriminant analysis and *k*-nearest neighbours are used to fit the model to the dataset to help separate the classes. Usually, these models first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification [1]. The best fitted model is chosen if its accuracy is the highest among the other models.

Missing Data

There are missing values in the dataset from Date of Construction and SubType. We approach this with mean/mode imputation which fill in the missing values with estimated ones. This avoids any important information being lost. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values [2].

Exploratory Data Analysis

Data Manipulation

In order to make the data more organised and readable, we create new variables as shown in Table 3. PD and PLA, and ELE and ELEDOO are combined since these variables are similar in tasks.

Table 3: New Variables

Combined variables	New variable
PD(Painter/Decorator) + Plasterer (PLA)	Decorator (DEC)
Electric (ELE)+ Electric Door (ELEDOO)	ELE (Electrician)
Routine repairs: 20 days (20D) + 3 days (3D)	Routine repairs (20D)
Current year, 2021 − Date of Construction	Age (Age of property)

Data Visualisation

There are four types of properties, including Flat and House, accounting to 44% (6172) and 42% (5929) of the total properties respectively (Figure 1).

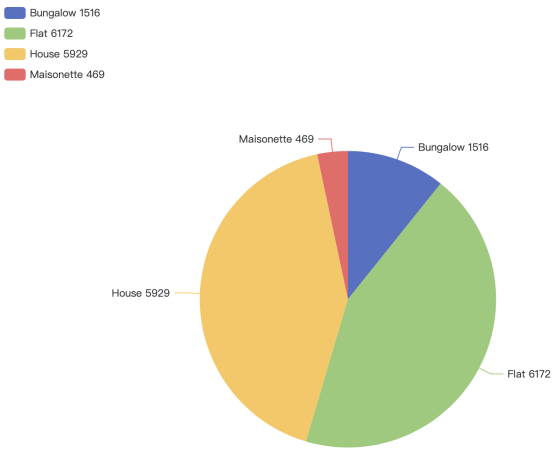

Figure 1: Total properties of each type.

Comparing with emergency and planned repairs, House has the highest number of routine repairs among the other types of properties (Figure 2).

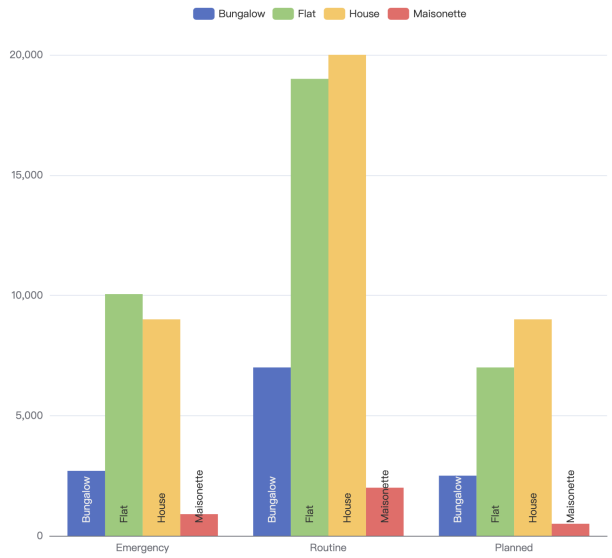

Figure 2: Total repairs in each time frame for each type of properties.

The repairs done were mostly for properties in PL5, and the total number of plumbing (PLU) and electric (ELE) repairs are the highest (Figure 3).

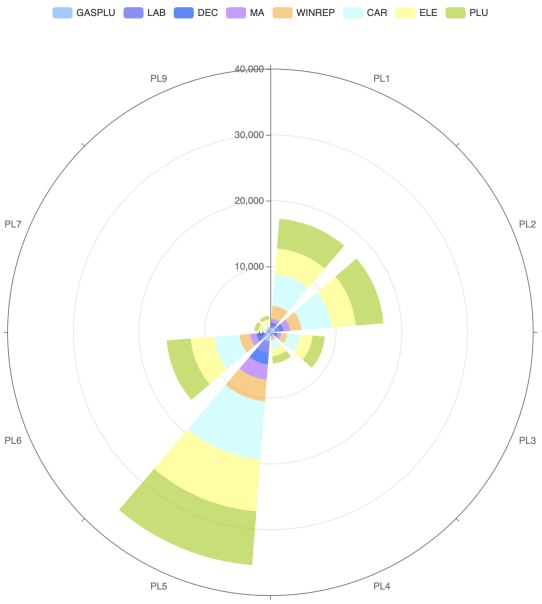

Figure 3: Total repairs for properties in each area.

Figure 4 shows that properties of age 60-80 years have the highest number of responsive repairs. As for >80 years, there are much less responsive repairs than expected. These properties may have been renovated, thus the need of these repairs is much lower.

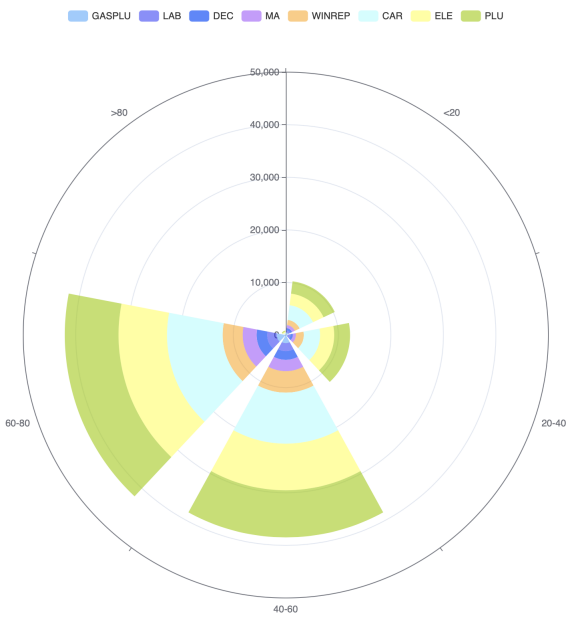

Figure 4: Total repairs for properties of each age group.

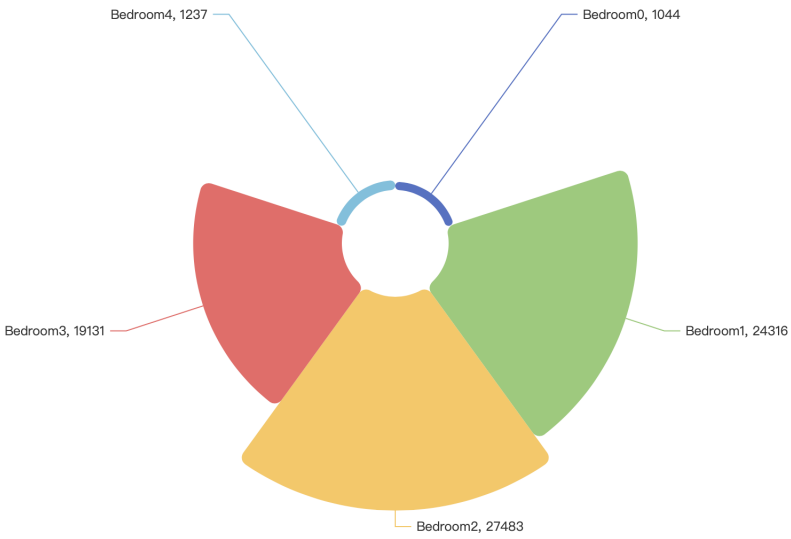

Figure 5: Total repairs for properties of different number of rooms.

Figure 5 shows that properties with two bedrooms have the highest number of total repairs, with more than 27,000 repairs, followed by those with one bedroom. Whereas, properties with zero bedroom (studio) have the lowest number of total repairs.

We also want to explore the distribution of each type of repair made within the three time frames in the Type and Postcode of properties. Figure 6 shows that routine repairs are highly requested in all four types of the properties and those in each of the nine districts. There are as many emergency repairs as planned repairs in most of the type of properties and districts.

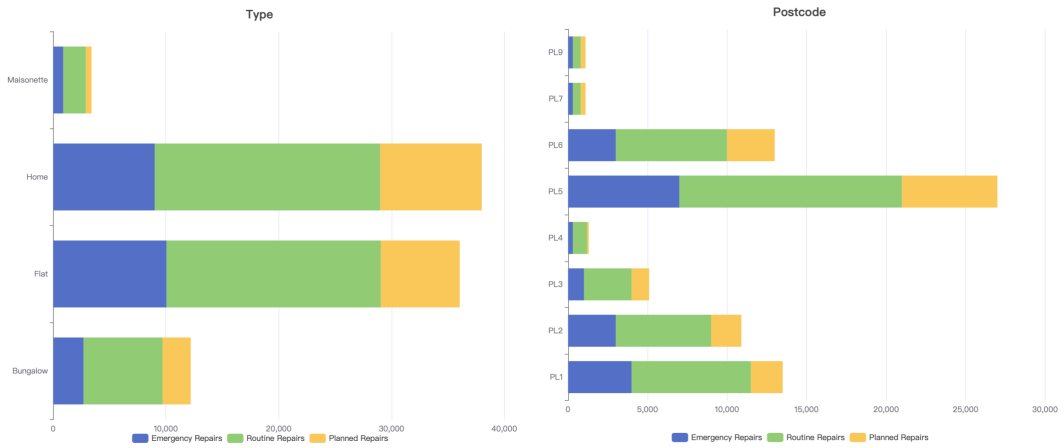

Figure 6: Comparison between Type and Postcode of the properties.

In the stacked bar plot (Figure 7), properties with only one tenant have more demand for repairs, mainly routine and emergency repairs.

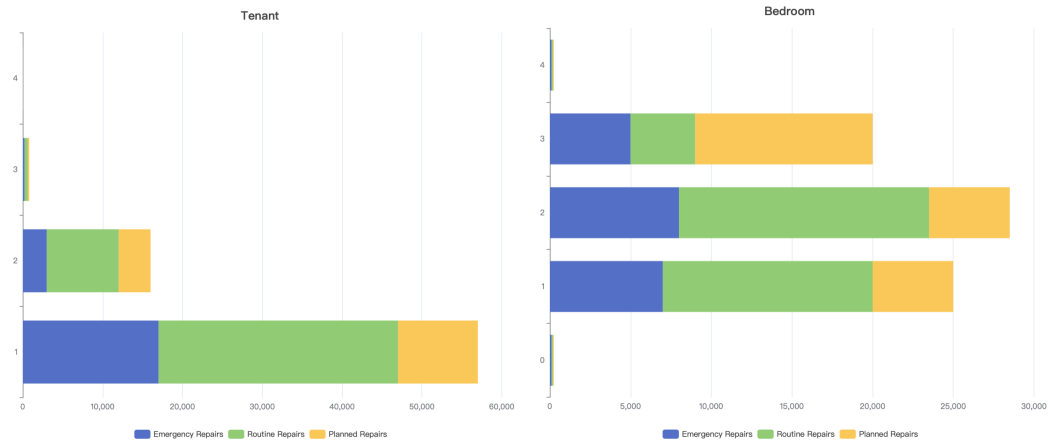

Figure 7: Comparison between Tenant and Bedroom of the properties.

Methods

Logistic Regression

Logistic regression models the probability of a certain class or an event existing. The event will be whether there is an emergency repair (yes, 1) or not (no, 0). It studies the relationship between a dichotomous dependant variable and the independent variables. The goal here is to find an equation that best predicts the probability of the event.

Logistic Function

Linear regression is not suitable for the investigation. This is due to the mismatch in the equation. We use logistic regression instead as it allows the probability to be mapped onto the entire real line. The equation is as follows:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{1i} + \cdots + \beta_{49} x_{49i} \tag{1}$$

- $p_i = \Pr(Y_i = 1 | x_{1i}, ..., x_{49i})$ is the probability of having an emergency repair for the i-th property, where $i \in [1, ..., n]$.
- $\alpha, \beta_1, ..., \beta_{49}$ are the model parameters,
- $x_1, ..., x_{16}$ are the number of routine and planned repairs of the 8 trades,
- categorical variables:
  - $x_{17}, ..., x_{20}$ - Bedrooms: 0, 1, 3, 4. If there are 2 bedrooms, $x_{17} = ... = x_{20} = 0$.
  - $x_{21}, ..., x_{24}$ - Age buckets: $< 20, > 80, 20 - 40, 40 - 60$. If the age is $60 - 80$, $x_{21} = ... = x_{24} = 0$.
  - $x_{25}, ..., x_{27}$ - Tenants: 2, 3, 4. If there is 1 tenant, $x_{25} = ... = x_{27} = 0$.
  - $x_{28}, ..., x_{30}$ - Types: Bungalow, House, Maisonette. If the type of property is a flat, $x_{28} = ... = x_{30} = 0$.
  - $x_{31}, ..., x_{42}$ - SubType: BISF, CORNISH UNIT, CROSSWALL, DORLONCO, EASIFORM, NO FINES, ORLIT, PASSIVHAUS, PREFAB, STAR, STONECRETE, TIMBER FRAMED. If the subtype of property is traditional, $x_{31} = ... = x_{42} = 0$.
  - $x_{43}, ..., x_{49}$ - Postcode: PL1, PL2, PL3, PL4, PL6, PL7, PL9. If the property is located in PL5, $x_{43} = ... = x_{49} = 0$.

In (1), there are some missing categories in each categorical variables (factors). **We set the reference category to a category that accumulates the highest in that corresponding factor**. By taking the exponential of both side of (1), we obtain the logistic function:

$$p_i = \frac{e^{\alpha+\beta_1 x_{1i}+...+\beta_{49} x_{49i}}}{1+e^{\alpha+\beta_1 x_{1i}+...+\beta_{49} x_{49i}}} \tag{2}$$

We then use maximum likelihood to fit the model (2) to the dataset. This will enable us to obtain an estimate of α and β. This idea can be defined using the likelihood function:

$$\ell(\alpha, \beta_1, ..., \beta_{49}) = \prod_{i=1} p_i^{Y_i} \prod_{i=1} (1-p_i)^{1-Y_i}, \quad Y_i = 1 \tag{3}$$

The parameter estimates are used to maximise (3). Machine learning helps to proceed this easily. Once we fit the model to the dataset, we can obtain the model parameter estimates.

Linear Discriminant Analysis (LDA)

LDA involves modelling the conditional distribution of Y given the predictor X.

Concept

There are two response classes, K, 0 and 1. We want to decide which property will most likely have an emergency repair. If we use only one predictor, $p = 1$ i.e. Age, it may not separate the classes well enough as shown in Figure 8 as there is an overlap. This is a one-dimensional case. The overlap can be minimised further if we include more predictors, as shown in Figure 9 and Figure 10. A p-dimensional case contains p predictors. It is harder to visualise the separation as the dimensions get bigger.

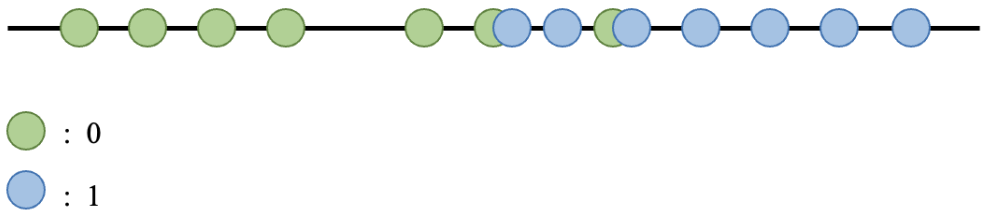

Figure 8: Distribution of the classes in a 1-D case.

LDA focuses on maximising the separability among the known classes by reducing a p-D graph to a 2-D graph in such a way that maximises the separability of the classes. It uses the information from both predictors to create a new axis and projects the data onto this new axis in a way to maximise the separation, as shown in Figure 11. This is known as the least squares method. The new axis is formed based on maximising the distance between the means of the classes and minimising the variation within each class resulting in well-separated classes.

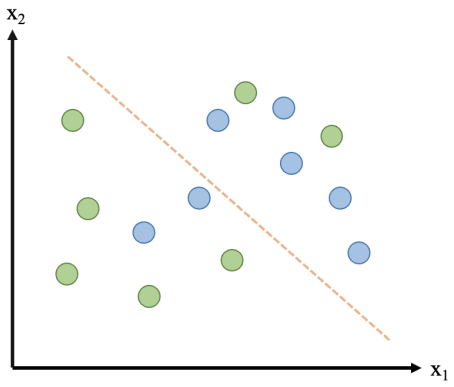

Figure 9: Distribution of the classes in a 2-D case. The orange dotted line separates the classes.

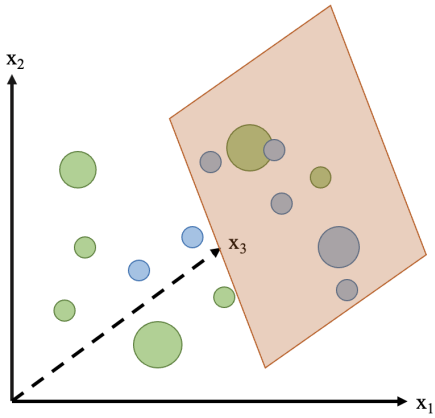

Figure 10: Distribution of the classes in a 3-D case. The orange plane separate the classes.

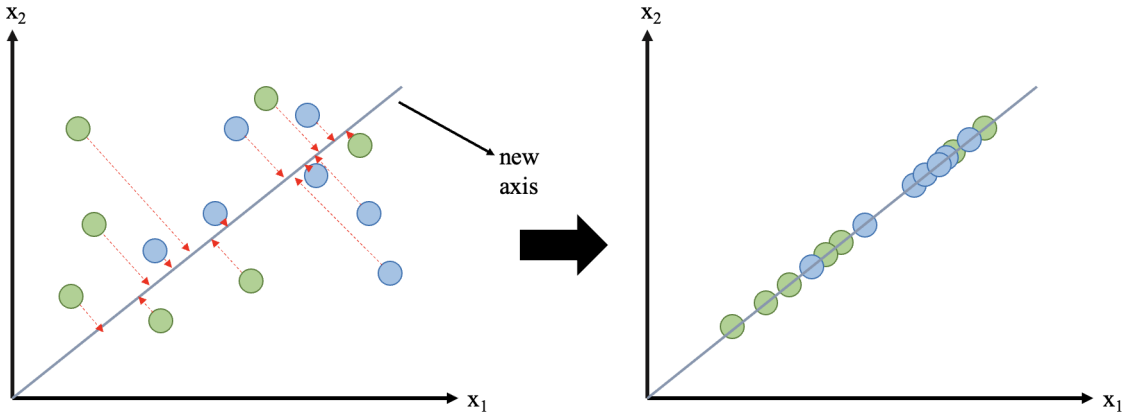

Figure 11: Reducing 2-D to 1-D graph by using least squares method.

Linear Discriminant Function

The linear discriminant function follows Bayes' approach. Recall that $K = 2$. Let $\Pr(X = x|K = k) = \pi_k$ be the prior probability that a randomly chosen observation comes from the k-th class, and $f_k(x) \equiv \Pr(X = x|Y = k)$. Then Bayes' theorem states that:

$$\Pr(Y = k|X = x) = p_k(X) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(k)} \tag{4}$$

Here $p_k(x)$ is the posterior probability that $X = x$ comes from the k-th class conditional on the predictor value. LDA estimates $f_k(x)$ to create a classifier that approximates the Bayes' classifier.

For $p = 1$:
Assume that $f_k(x)$ follows a normal distribution, and $\sigma_1^2 = ... = \sigma_K^2$ so we can simplify them to just be σ^2. μ_k and σ_k^2 denotes the mean and variance parameters for the k-th class. LDA will approximate the Bayes classifier and classifies to the class for which $\hat{\delta}_k(x)$ is largest by substituting estimates of π_k, μ_k and σ^2 into:

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k), \tag{5}$$

where the Bayes' decision boundary corresponds to the point,

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}, \tag{6}$$

to obtain the following linear decision boundary:

$$\hat{\delta_k}(x) = x\frac{\hat{\mu_k}}{\hat{\sigma^2}} - \frac{\hat{\mu_k^2}}{2\hat{\sigma^2}} + \log(\hat{\pi_k}) \tag{7}$$

The estimates are calculated as follows:

$$\hat{\mu_k} = \frac{1}{n_k}\sum_{i:y_i=k} x_i, \tag{8}$$

$$\hat{\sigma^2} = \frac{1}{n-K}\sum_{k=1}^{K=1}\sum_{i:y_i=k}(x_i - \hat{\mu_k})^2. \tag{9}$$

$$\hat{\pi_k} = \frac{n_k}{n} \tag{10}$$

where n is the total number of training observations and n_k is that in the k-th class. Note that σ^2 is the weighted average of the sample variances for each of the K classes. To implement LDA, we begin by estimating π_k, μ_k and σ^2. We then compute the decision boundary, that results from assigning an observation to the class for which (7) is largest [1].

For $p > 1$:
We assume that $X = (X_1, ..., X_p)$, and each predictor follows a one-dimensional normal distribution. The discriminant function is simply the vector/matrix version of (7):

$$\delta_k(x) = \frac{x^T\mu_k}{\sum} - \frac{\mu_k^T\mu_k}{2\sum} + \log(\pi_k) \tag{11}$$

Here, μ_k is a class-specific mean vector, $\sum$ is a covariance matrix that is similar to all K classes and π_k is the prior probability that an observation belongs to the k-th class [1]. We will need to estimate $\mu_1, ..., \mu_K$, $\pi_1, ..., \pi_K$ and $\sum$ by using formulas similar to (8) and (9). Again, LDA substitutes these estimates into (11) and classifies to the class for which $\hat{\delta}_k(x)$ is largest.

k-Nearest Neighbours

The *k*-nearest neighbours (*k*-NN) is a non-parametric algorithm used for classification. The model is distributed from the distance of data, which means it is a supervised classification based on the train dataset. By using this approach, properties can be separated into several classes to predict the probability of emergency repairs of a new sample.

Concept

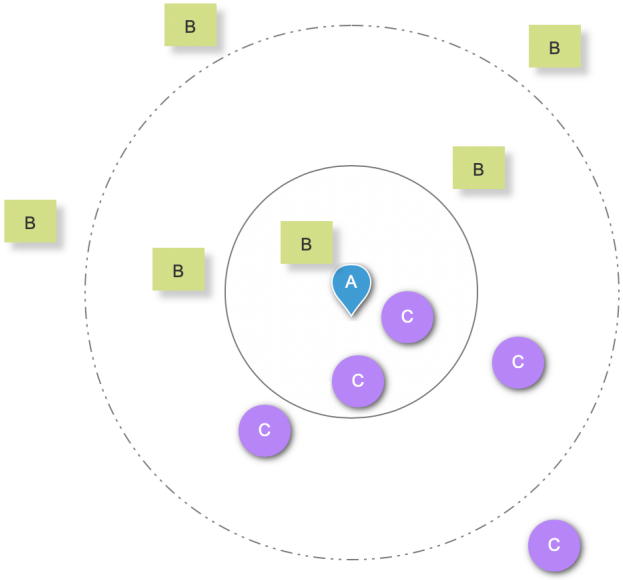

Figure 12: Example of *k*-NN.

Figure 12 shows how *k*-NN works. The value of k is the number of the nearest neighbours of the new samples. The central A (blue) is the sample which should be classified to other types, either class B (green) or class C (purple). If $k = 3$, according to the distances between A and other variables, the test sample A should be classified into class C, as the smallest circle has two purple circles and one green squares (2>1). If $k = 7$, according to the distances between A and other variables, the test sample A should be classified into class C, as the outer circle has four purple circles and three green squares (4>3).

In order to avoid the over-fitting problem, the value of k should be selected through cross-validation. We assume that the predictor $X = (X_1, ..., X_p)$. The distance is measured by distance metrics:

$$d(X_1, X_2) = (\sum_{i=1}^{n} |X_{1i} - X_{2i}|^p)^{\frac{1}{p}}, n \text{ is the dimensionality of the space.} \quad (12)$$

When $p = 1$, we obtain the Manhattan distance. When $p = 2$, we obtain the Euclidean distance.

Decision Tree

The classification tree is constructed with the binary decision structure. This means that the root and all subsequent branches can grow only two new nodes. The goal of the classification tree is to make an optimal selection of whether a property has an emergency repair.

Concept

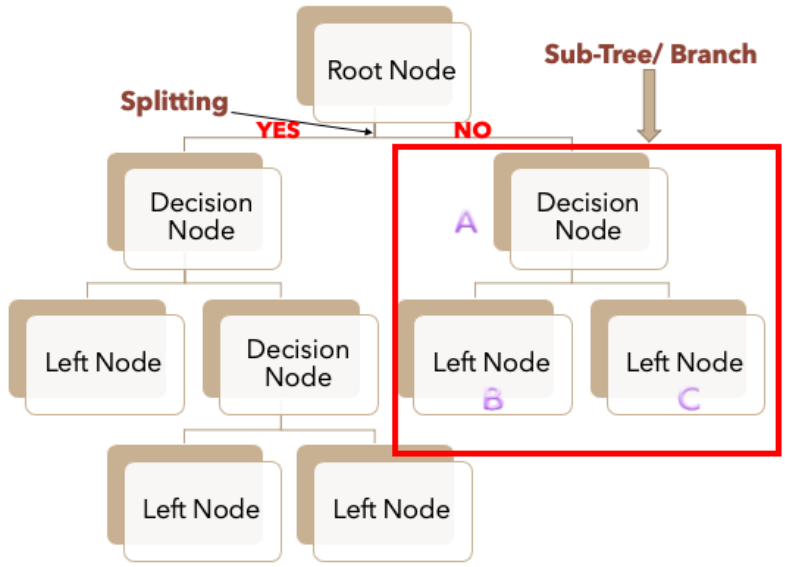

Figure 13: Significant terms in the classification tree.

Each node in the tree represents an attribute test (Figure 13). The decision process starts at the root node and follows the branch until it reaches the leaf node. This process is repeated until all instances in a node are in the same class, or further splitting fails to improve the prediction.

The first node in the tree, which represents the entire population or sample, is further split into two or more homogeneous sets. A branch represents a test outcome (either yes or no) and connects to the next node or leaf. When a sub-node splits into further child nodes, it is called the decision node. The process of deleting the child nodes of the decision node is called pruning. At the end of the tree are the leaf nodes (external nodes), which represent the predicted results.

When a node splits 40/60, it is 100% impure, and when all the data belongs to a single class, it is 100% pure. Methods of entropy or Gini index is used to evaluate the quality of classification tree splitting and determine that it will produce the best outcome. The entropy calculation formula is as follows: [1]

$$E(X) = -\sum_{i=k}^{k} \hat{p}_{mk} \log(\hat{p}_{mk}) \tag{13}$$

where $\hat{p}_{mk}$ represents the proportion of training observations from the k-th class in the m-th region. K represents either "1 (Yes)" or "0 (No)". The higher the **entropy** represents a more rare outcome from the information [5].

The Gini index formula is given by:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{14}$$

When a region m contains most of the data from a single class K, the Gini Index value will be small [1]. The higher the **Gini index**, the higher the degree of inequality and the greater the heterogeneity.

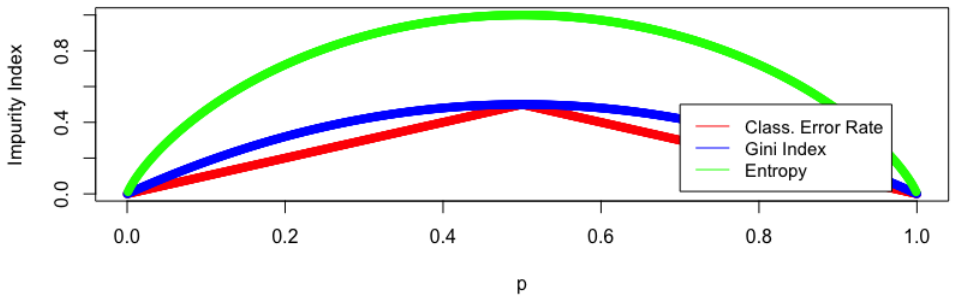

Figure 14: Comparing the effectiveness method for nodes impurity.

Figure 14 shows that the **Classification Error Rate** is insensitive to the growth of the tree. In practical application, the other two methods are preferable.

$$E = 1 - \max_k(\hat{p}_{mk}) \tag{15}$$

The classification tree tends to over-fit because at each node, it makes a decision in a subset of all the features, and when it reaches the final decision, it is a complex decision chain. Pruning is a method to avoid classification tree over-fitting [10].

Random Forest

Random Forest is rather a bootstrap aggregation method used in model prediction. It applies the method of bagging to measure the relative importance of each variable on the prediction. The aim is to predict the value of the responsive feature and whether there is an emergency repair.

Concept

Random forest uses the bootstrap method to create multiple bootstrap datasets, as shown in Figure 15. Bootstrapping is a statistical process of re-sampling a training set where 100% of the observed data was replaced with samples. This model randomly selects a predefined number of features as candidates to create n classification trees with different bootstrapped data sets. Bootstrapping ensures that each classification

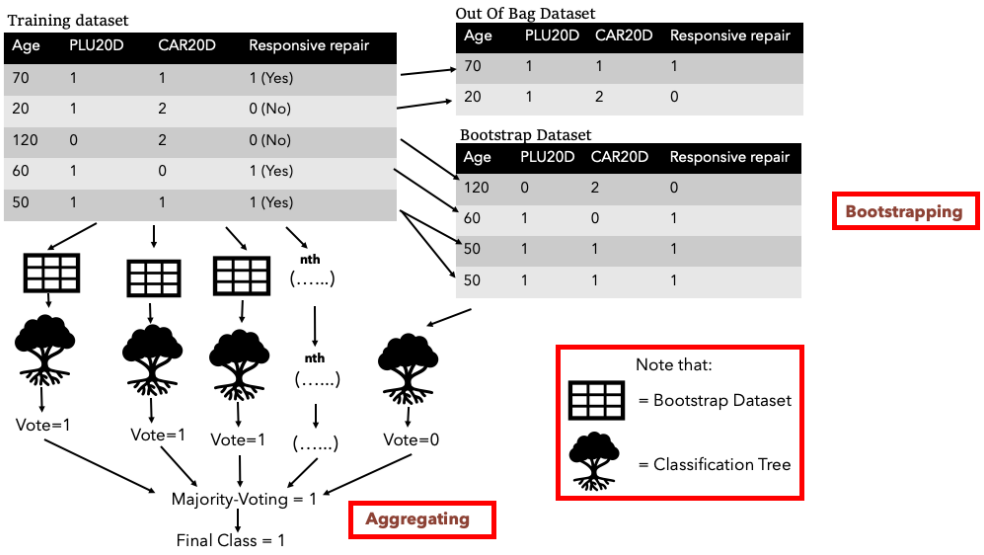

Figure 15: How the model works

tree in the random forest is unique, which reduces the population variance of the random forest classifiers.

The random forest classifier aggregates the decisions of individual trees and makes the final decision by majority vote. The prediction for a classification problem is:
$f(x)$ = majority vote of all predicted classes over n classification trees

Any sample that is not selected for the bootstrap dataset (approximately 1/3 of the training dataset) is placed in a separate data-set, called the Out-of-Bag dataset [12]. This dataset is used for calculating the importance of a specific variable and estimate the random forest's goodness-of-fit.

Results

Logistic Regression

Interpretation of the Result

After fitting the model to the dataset, we perform backward elimination to remove insignificant numerical variables to p and likelihood ratio test to determine which categorical variable is not needed, the result obtained is shown in Table 4 and Table 5.

Numerical variables:
Recall that $\boldsymbol{Y_i}$ **indicates of having an emergency repair**. In Table 4, $\hat{\alpha}$ is the log odds of Y_i given that all other variables are 0. However, this would not make sense realistically as this is an example of extrapolation. $\hat{\beta}_i$ is the change in log-odds of Y_i for every one repair increase in x_i, given that the status of other variables are fixed. Similarly, it also represents the log odds ratio of Y_i between the number of two sets of repairs made within a time period. Take $\hat{\beta}_1$ for instance, 0.241 represents the log odds ratio of Y_i between $(k+1)$ and k carpenter repairs made within 20 days (CAR20D),

Table 4: Results for the numerical variables obtained from the logistic regression model.

Numerical variable	Parameter estimate	95% confidence interval
-	$\hat{\alpha}$ = -0.00334	(-0.111, 0.127)
CAR20D, x_1	$\hat{\beta_1}$ = 0.241	(0.184, 0.300)
CAR60D, x_2	$\hat{\beta_2}$ = 0.194	(0.0997, 0.290)
ELE20D, x_3	$\hat{\beta_3}$ = 0.201	(0.160, 0.243)
ELE60D,x_4	$\hat{\beta_4}$ = 0.199	(0.144, 0.256)
GASPLU20D, x_5	$\hat{\beta_9}$ = -0.263	(0.0213, 0.194)
MA20D, x_9	$\hat{\beta_{10}}$ = 0.107	(0.0404, 0.218)
MA60D, x_{10}	$\hat{\beta_{11}}$ = 0.129	(0.0555, 0.309)
DEC20D, x_{11}	$\hat{\beta_{13}}$ = 0.181	(0.212, 0.286)
PLU20D, x_{13}	$\hat{\beta_{14}}$ = 0.248	(0.147, 0.418)
PLU60D, x_{14}	$\hat{\beta_{13}}$ = 0.280	(0.212, 0.286)
WINREP20D, x_{15}	$\hat{\beta_{14}}$ = 0.156	(0.147, 0.418)
WINREP60D, x_{16}	$\hat{\beta_{14}}$ = 0.194	(0.147, 0.418)

where $k = 1, ...$ The odds ratio will simply be exp(0.241)=1.27.

$\hat{\boldsymbol{\beta}}_{\mathbf{5}}$ **is negative**. This indicates that if we increase the number of gas plumbing repairs within 20 days (GASPLU20D), p_i will decrease given that the other variables are at a fixed status. As for the **positive parameter estimates**, i.e. $\hat{\beta_{16}}$, increasing the number of window repairs within 60 days (WINREP60D) will increase p_i given that other variables are at a fixed status.

The 95% confidence interval does not contain 0, indicating that we can rule out the possibility that these parameters are 0. It also shows that the fitted model contain the necessary numerical variables that are statistically significant.

Categorical variables:
The estimates shown in Table 5 is the log odds ratio of Y_i between a category in its factor and its reference category, given that other variables are fixed. Therefore, the log odds ratio of Y_i between three and two bedrooms is 0.0647. The corresponding odds ratio is then exp(0.0647)=1.07. Here, it seems that the odds that properties with three bedrooms are 1.07 times more likely to have an emergency repair than those with two bedrooms. As for those with one bedroom, the odds ratio is exp(-0.133)=0.875. To interpret this value in an easier way, we invert it. So the odds ratio of having an emergency repair for properties with two bedrooms is 1/0.875=1.14 times more that those with one bedroom. A summary of the log odds ratio for the rest of the categorical variables can be seen in Table 5.

Some of the 95% confidence interval in Table 5 does not contain 0. This just indicates that we can rule out the possibility that the log odds ratio of having an emergency repair are significantly different than 0. We cannot rule out such possibility otherwise.

Table 5: Results for the categorical variables obtained from the logistic regression model. There are two confidence intervals containing no upper bound. This simply means that the interval is wider and may lack precision due to high dispersion within the corresponding variable's data.

Categorical variable	Parameter estimate	95% confidence interval
Bedrooms: 0, x_{17}	$\hat{\beta_{17}}$ = -0.333	(-0.627, -0.0363)
1, x_{18}	$\hat{\beta_{18}}$ = -0.133	(-0.247, -0.0189)
3, x_{19}	$\hat{\beta_{19}}$ = 0.0647	(-0.0428, 0.172)
4, x_{20}	$\hat{\beta_{20}}$ = 0.309	(0.0332, 0.590)
Age: < 20, x_{21}	$\hat{\beta_{21}}$ = 0.0352	(-0.114, 0.184)
> 80, x_{22}	$\hat{\beta_{22}}$ = 0.329	(-0.0906, 0.766)
20-40, x_{23}	$\hat{\beta_{23}}$ = 0.0802	(-0.0642, 0.225)
40-60, x_{24}	$\hat{\beta_{24}}$ = 0.166	(0.0487, 0.284)
Tenants: 2, x_{25}	$\hat{\beta_{25}}$ = 0.191	(0.0466, 0.337)
3, x_{26}	$\hat{\beta_{26}}$ = 0.649	(0.231, 1.09)
4, x_{27}	$\hat{\beta_{27}}$ = 11.0	(-6.39, **NA**)
Type: Bungalow, x_{28}	$\hat{\beta_{28}}$ = 0.534	(0.379, 0.690)
House, x_{29}	$\hat{\beta_{29}}$ = -0.369	(-0.486, -0.252)
Maisonette, x_{30}	$\hat{\beta_{30}}$ = 0.105	(-0.139, 0.354)
SubType: BISF, x_{31}	$\hat{\beta_{31}}$ = 0.131	(-0.132, 0.398)
CORNISH UNIT, x_{32}	$\hat{\beta_{32}}$ = -0.0662	(-0.209, 0.0770)
CROSSWALL, x_{33}	$\hat{\beta_{33}}$ = -0.0765	(-0.366, 0.217)
DORLONCO, x_{34}	$\hat{\beta_{34}}$ = 0.605	(-1.78, 3.69)
EASIFORM, x_{35}	$\hat{\beta_{35}}$ = 0.0795	(-0.0665, 0.226)
NO FINES, x_{36}	$\hat{\beta_{36}}$ = 0.00624	(-0.165, 0.179)
ORLIT, x_{37}	$\hat{\beta_{37}}$ = 0.178	(-0.356, 0.738)
PASSIVHAUS, x_{38}	$\hat{\beta_{38}}$ = -0.656	(-1.26, -0.0576)
PREFAB, x_{39}	$\hat{\beta_{39}}$ = 11.3	(-54.9, **NA**)
STAR, x_{40}	$\hat{\beta_{40}}$ = 0.312	(0.0463, 0.585)
STONECRETE, x_{41}	$\hat{\beta_{41}}$ = -2.11	(-5.20, -0.124)
TIMBERFRAMED, x_{42}	$\hat{\beta_{42}}$ = -0.309	(-0.517, -0.0995)

Logistic Function

By substituting the parameter estimates into (2), we obtain the following:

$$\hat{p_i} = \frac{e^{-0.00334+0.241x_{1i}+...-0.309x_{42i}}}{1+e^{-0.00334+0.241x_{1i}+...-0.309x_{42i}}}$$

$x_{43i}, ..., x_{49i}$ is not in the equation anymore as the variable Postcode was proven not needed in the model. We can then calculate p_i i.e. the probability of having an emergency repair given that five carpenter repairs has been completed within 20 days and that the property is of two bedrooms, aged 20-40, has one tenant and is a traditional flat is:

$$\hat{p_i} = \frac{e^{-0.00334+0.241(5)+0.0802(1)}}{1+e^{-0.00334+0.241(5)+0.0802(1)}} = 0.783$$

We also want to determine how well did the logistic regression model predict the probability and overcome this classification problem.

Predicted Accuracy

Table 6: Confusion matrix based on the training set.
TN: True negative; FN: False negative; FP: False positive; TP: True positive

Predicted/Actual classes	0	1
0	TN=259	FN=194
1	FP=695	TP=1,669

Now, we **fit the model to the training set (80% of the dataset)**. In table 6, we see that the model correctly predicted that 1,669 properties will have an emergency repair, and that 259 properties will not have an emergency repair. Therefore, the predicted accuracy is $(1,669+259)/(1,669+259+695+194) \times 100 = 68.4\%$

We can confirm this accuracy by looking at the receiver operating characteristics (ROC) curve in Figure 16. AOC is the area under the curve (blue region). The greater the area, the higher the accuracy. Specificity and sensitivity represents the true negative rate and true positive rate respectively:

$$\text{Sensitivity} = \text{True positive rate} = \frac{\text{TP}}{\text{Actual yes (1)}}$$

$$\text{Specificity} = \text{True negative rate} = \frac{\text{TN}}{\text{Actual no (0)}}$$

A perfect ROC curve approaches closely to the top left corner, indicating a high true positive rate and a low false positive rate. The dotted line represents the ”no information” classifier. We can see that the area under the curve equates to 0.688 which is decent.
Figure 17 shows a plot of the predicted probabilities based on the training set along with the rank of each sample, from low to high probability. Most of the properties that has an emergency repair are predicted to have a high probability of Y_i, and those with no emergency repair are predicted to have a low probability of Y_i. The graph also shows an (almost) S-curve.
We can also extract the highest and lowest probability of Y_i and its index to determine the property associated to it. This model predicts that properties located in PL2 3JH and PL1 5JT obtains the highest probability of having an emergency repair, while PL2 2PN and PL1 4HL obtains the lowest probability.

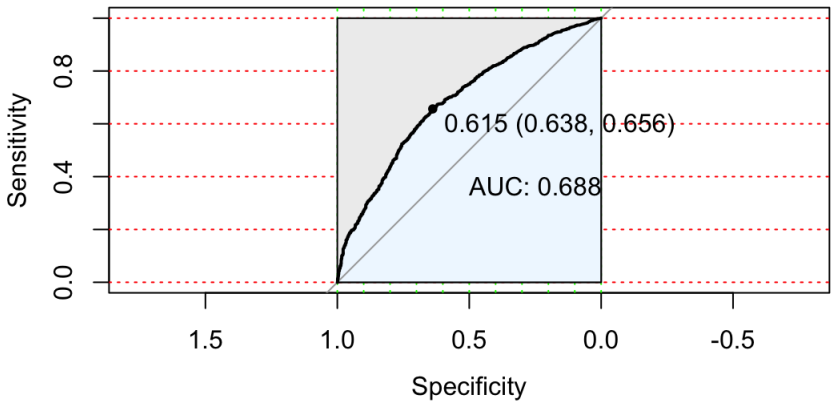

Figure 16: ROC curve for the logistic regression model on the training set.

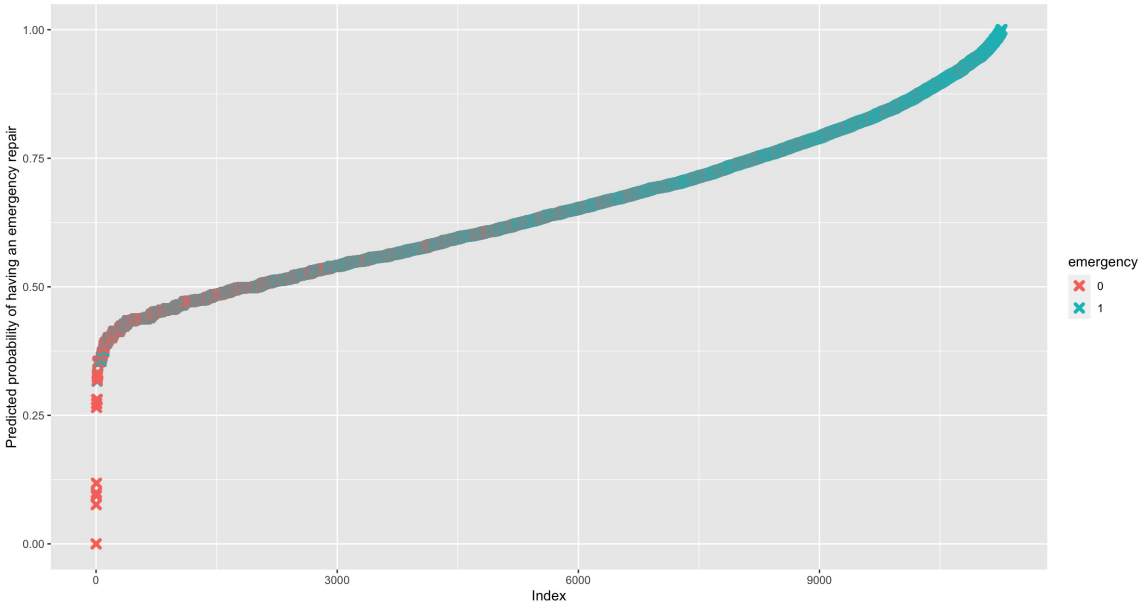

Figure 17: A plot showing the predicted probabilities that each property has an emergency repair along with the other variable statuses.

Linear Discriminant Analysis (LDA)

Interpretation of the Result

Based on Table 7, we see that $\hat{\pi_1} = 0.346$ and $\hat{\pi_2} = 0.654$. Specifically, 34.6% of the training observations correspond to properties to which there is no emergency repair.

Table 7: The estimated prior probabilities, $\hat{\pi_1}$ and $\hat{\pi_2}$ within each class.

Prior probabilities of groups:		
Group	0	1
Prior probability	0.346	0.654

The average shown in Table 9 could suggest that i.e. the number of window repairs, x_{16}, have a slightly greater influence on having an emergency repair than on not having one.

The values in Table 10 contributes to the linear discriminant function, which is $0.269 \times x_1 + 0.183 \times x_2 + ... - 0.479 \times x_{42}$. If it is large, the property will have an emergency repair and if it is small, it will predict otherwise.

Predicted Accuracy

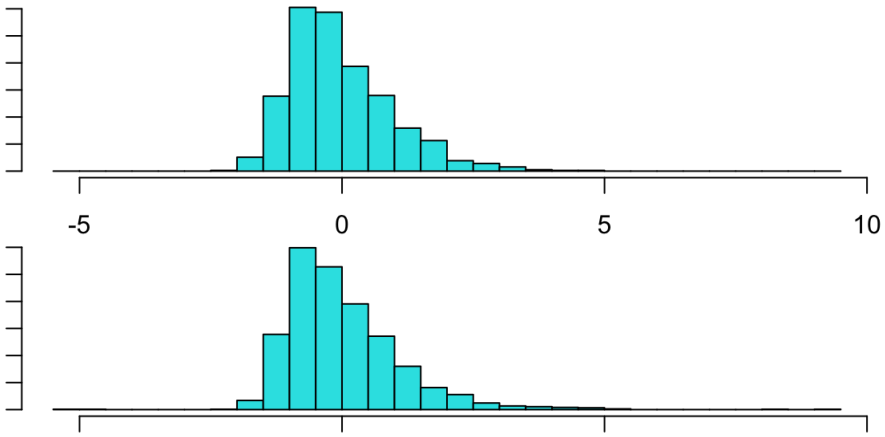

Figure 18: Plots of the linear discriminants obtained by computing the linear discriminant function for each of the training observations. Top: Class 0; Bottom: Class 1.

In Figure 18, we see that the classes are not well-separated as they overlap. Again, we check the performance of the model by observing the confusion matrix in Table 8.

The model correctly predicted that 1,753 properties will have an emergency repair, and that 159 properties will not have an emergency repair. Therefore, the predicted accuracy is $(1,753+159)/(1,753+159+795+110) \times 100 = 67.9\%$. **LDA underperforms compared to the logistic regression model**. The accuracy of the predicted probability can be checked by plotting the ROC curve, as seen in Figure 19. The area under the curve is 0.684, which is slightly less than that obtained for the logistic regression model.

Table 8: Confusion matrix based on the training set.

Predicted/Actual classes	0	1
0	TN=159	FN=110
1	FP=795	TP=1753

Table 9: The estimated average of each predictor within each class.

Group means:		
Predictor	Class: 0	1
CAR20D, x_1	0.266	0.558
CAR60D, x_2	0.102	0.218
ELE20D, x_3	0.544	0.953
ELE60D, x_4	0.277	0.520
GASPLU20D, x_5	0.121	0.144
MA20D, x_9	0.133	0.229
MA60D, x_{10}	0.116	0.199
DEC20D, x_{11}	0.0598	0.134
PLU20D, x_{13}	0.636	1.20
PLU60D, x_{14}	0.0487	0.126
WINREP20D, x_{15}	0.283	0.481
WINREO60D, x_{16}	0.0408	0.0640
Bedrooms: 0, x_{17}	0.0190	0.0157
1, x_{18}	0.284	0.315
3, x_{19}	0.310	0.286
4, x_{20}	0.0185	0.0221
Age: < 20, x_{21}	0.113	0.0923
> 80, x_{22}	0.00718	0.00868
20-40, x_{23}	0.107	0.115
40-60, x_{24}	0.339	0.377
Tenants: 2, x_{25}	0.118	0.164
3, x_{26}	0.00692	0.0167
4, x_{27}	0.00	0.000407
Type: Bungalow, x_{28}	0.0692	0.129
House, x_{29}	0.471	0.392
Maisonette, x_{30}	0.0254	0.0377
SubType: BISF, x_{31}	0.0223	0.0240
CORNISH UNIT, x_{32}	0.121	0.101
CROSSWALL, x_{33}	0.0195	0.0149
DORLONCO, x_{34}	0.00	0.000271
EASIFORM, x_{35}	0.105	0.107
NO FINES, x_{36}	0.0531	0.0668
ORLIT, x_{37}	0.00462	0.00475
PASSIVHAUS, x_{38}	0.00462	0.00258
PREFAB, x_{39}	0.00	0.000136
STAR, x_{40}	0.0205	0.0294
STONECRETE, x_{41}	0.00103	0.00
TIMBERFRAMED, x_{42}	0.0382	0.0419

Table 10: The coefficients of linear discriminants used to form the LDA decision rule.

Predictor	Coefficients of linear discriminants
CAR20D, x_1	0.269
CAR60D, x_2	0.183
ELE20D, x_3	0.241
ELE60D, x_4	0.192
GASPLU20D, x_5	-0.383
MA20D, x_9	0.139
MA60D, x_{10}	0.183
DEC20D, x_{11}	0.178
PLU20D, x_{13}	0.309
PLU60D, x_{14}	0.325
WINREP20D, x_{15}	0.162
WINREP60D, x_{16}	0.280
Bedrooms: 0, x_{17}	-0.350
1, x_{18}	-0.194
3, x_{19}	0.132
4, x_{20}	0.598
Age: < 20, x_{21}	0.0188
> 80, x_{22}	0.470
20-40, x_{23}	0.223
40-60, x_{24}	0.322
Tenants: 2, x_{25}	0.299
3, x_{26}	0.854
4, x_{27}	1.42
Type: Bungalow, x_{28}	0.721
House, x_{29}	-0.584
Maisonette, x_{30}	0.112
SubType: BISF, x_{31}	0.479
CORNISH UNIT, x_{32}	-0.0988
CROSSWALL, x_{33}	-0.295
DORLONCO, x_{34}	3.17
EASIFORM, x_{35}	0.256
NO FINES, x_{36}	0.0365
ORLIT, x_{37}	0.434
PASSIVHAUS, x_{38}	-0.563
PREFAB, x_{39}	1.84
STAR, x_{40}	0.579
STONECRETE, x_{41}	-3.52
TIMBERFRAMED, x_{42}	-0.479

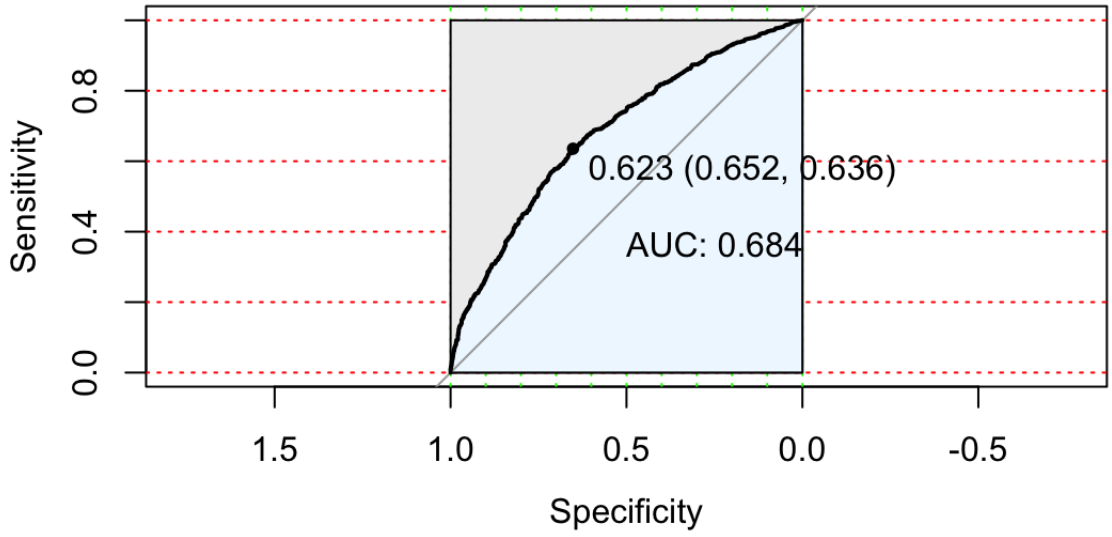

Figure 19: ROC curve for the LDA classifier on the training set.

Again, we can extract the highest and lowest predicted probability of a having an emergency repair. Properties in PL2 2PN and PL1 4HL obtains the highest probability while PL3 6SW and PL1 5HW obtains the lowest probability. There is a difference in these probabilities where the highest probability here is instead the lowest probability predicted by the logistic regression model.

k-Nearest Neighbours (*k*-NN)

Interpretation of Result

Since the *k*-NN algorithm needs to calculate the distance between different samples, the data types of variables are all needed to change into numeric. Different variables have different scales, which means the range and unit of variables should be normalised in order to simplify the calculation. We use min-max normalisation to uniform the range of values, which formulates to:

$$X_{\text{new}} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{16}$$

where $X_{\text{new}} \in [0, 1]$.
For the binomial variables, like y, we keep the values of it. For the categorical variables which has more than two classes (discrete variables), if the variables work with ordered-factors, we covert them into normal factors.

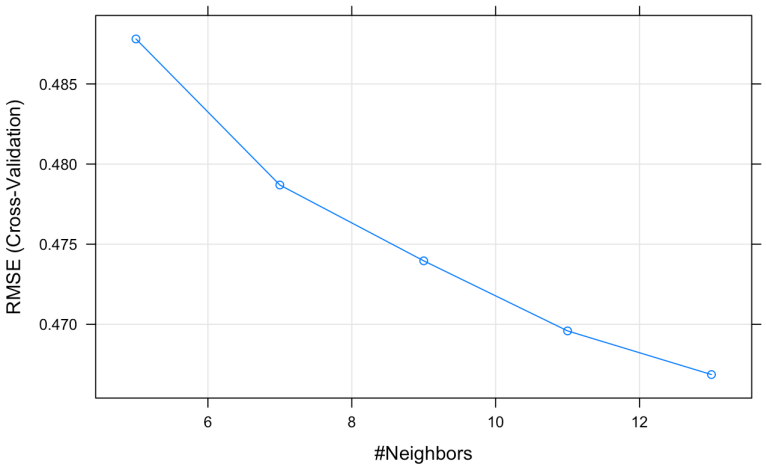

Figure 20: Plot of RMSE of the Cross-Validation

RMSE is used to select the optimal model using the smallest value (Figure 20). The final value used for the model is k = 13.
The *k*-NN algorithm can provide a rough classification of samples. The disadvantage of this algorithm is that it cannot have the ranking of variables about the importance. That is to say, *k*-NN only can predict the data, but not determine which variables have more effects on the results. In general, we use random forest as an additional method because it can provide more detailed information about the variables.

Predicted Accuracy

Table 11: Confusion matrix based on the training set.

Predicted/Actual classes	0	1
0	TN=176	FN=853
1	FP=205	TP=1703

This model correctly predicted that 1,703 properties will have an emergency repair, and that 176 properties will not have an emergency repair. Thus, the accuracy for *k*-NN is $(1,703 + 176)/(1,703 + 176 + 205 + 853) \times 100 = 64.1\%$.

Decision Tree

We separate the dataset into 80% training set and 20% testing set. We use the following features to build the classification tree:

- Numerical variables:
  - $x_1, ..., x_{16}$ are the number of routine and planned repairs of each of the 8 trades.
- Categorical variables:

- Bedrooms: 0, 1, 3, 4
- Age buckets: $< 20, > 80, 20 - 40, 40 - 60$
- Tenants: 2, 3, 4
- Types: Bungalow, House, Maisonette
- SubType: BISF, CORNISH UNIT, CROSSWALL, DORLONCO, EASIFORM, NO FINES, ORLIT, PASSIVHAUS, PREFAB, STAR, STONECRETE, TIMBER FRAMED
- Postcode: PL1, PL2, PL3, PL4, PL6, PL7, PL9.

We set the reference category to a category that accumulates the highest in that corresponding factor.

Fit Process

We use 10-fold cross-validation to find out how the error in the tree varies with the size of the tree.

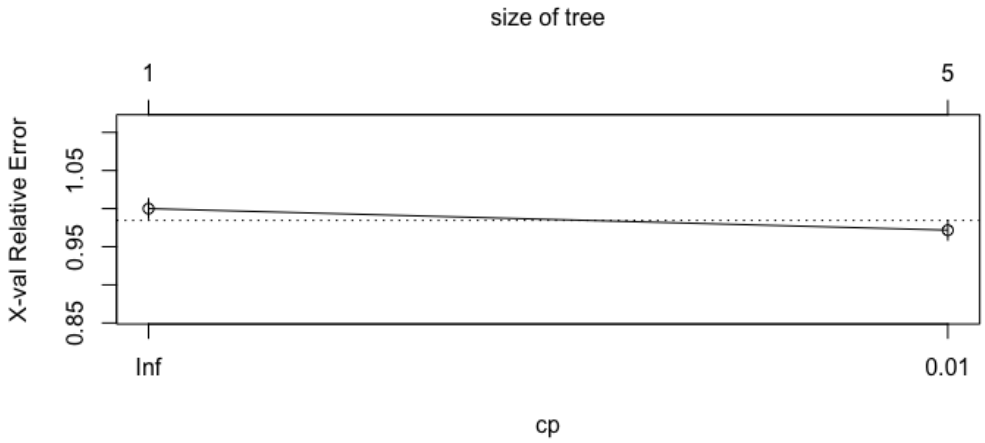

Figure 21: 10-fold Cross-Validation

Table 12: Complexity Table:
Number of split (nsplit); 10-fold Cross-Validated Error Rate (xerror); Standard Error (xstd)

Complexity parameters	nsplit	relative error	xerror	xstd
0.011	0	1.00	1.0000	0.012994
0.010	4	0.94	0.97166	0.012904

10-fold cross-validation creates 10 random subsets of the training data, using one portion of them as a test set. It builds a classification tree for the remaining nine parts and evaluates the tree using the test portion.

The tree that produces the lowest cross-validation error rate (xerror) is selected as the tree that most fits the data [16]. In this case, this minimal 10-fold cross-validated error rate is 0.97166 (Table 12), and the tree has four splits.
As shown in Table 12, there is a relatively small error at a complexity parameter (cp) of 0.01, indicating that the optimal pruning occurs when the tree has four splits (Figure 21).

Interpretation of the Result

Figure 22 demonstrates the pruned classification tree used to classify emergency repairs as Yes (1) or No (0) based on the "features" of the property, such as Type (House), x_{13} (PLU20D), x_3 (ELE20D) and x_{15} (WINREP20D).

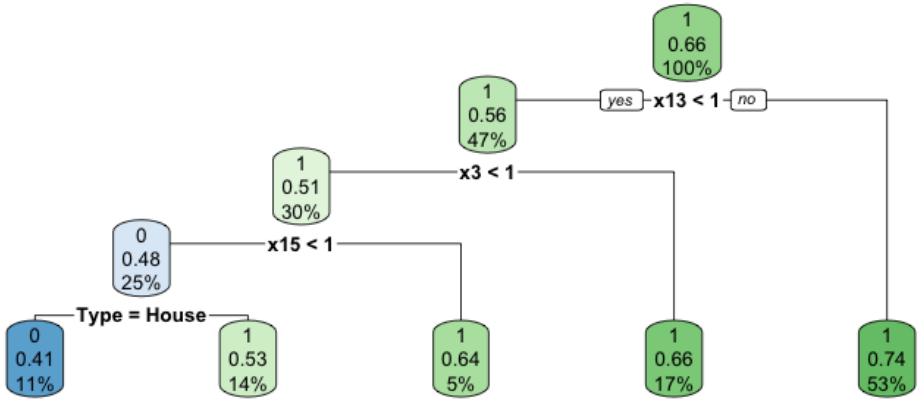

Figure 22: Pruned Classification Tree

The overall probability of whether properties have an emergency repair is at the root node. It shows that the proportion of properties that do not have an emergency repair is 66% (Figure 22). This node asks whether x_{13} is less than one. If so, we look at the root's left child node; otherwise, on the root's right child node.

To the root's right child node, 53% of the properties have $x_{13} > 1$. At the $x_{13} > 1$ node, the probability of properties with no emergency repair is 74%. So, the overall terminal node of this bucket ends with not having an emergency repair.

To the root's left child node, 47% of the properties have the feature $x_{13} < 1$. At this node, the probability of properties with no emergency repair is 56%. The node splits again between whether x_3 is less than one. If so, we move downwards towards the left; otherwise, towards the right.

The interpretation seems like, "If a property has the features $(x_{13} < 1)$, $(x_3 < 1)$, $(x_{15} < 1)$ and (Type: House), then $(y = 0)$". It means that properties having these kinds of features would not have an emergency repair. We can continue to learn which features affect the likelihood of emergency repair.

Predicted Accuracy

As shown in table 13, this model correctly predicted that 1,722 properties will have an emergency repair, and that 166 properties will not have an emergency repair. Thus, the accuracy of the pruned tree is $(1,722+166)/(1,772+166+805+125)\times 100 = 67.0\%$.

Table 13: Confusion matrix based on the training set.

Predicted/Actual classes	0	1
0	TN=166	FN=805
1	FP=125	TP=1722

So far, the accuracy of the decision tree is the lowest and the logistic regression model is the highest among the four models discussed.

Random Forest

Fit Process

We separate the dataset into 80% training set and 20% testing set. The features that we use to build the random forest are those used for the decision tree. Again, we set the reference category to a category that accumulates the highest in that corresponding factor.
We tune hyper-parameters to reduce the error rate in the model including m_{try} (number of variables randomly sampled as candidates at each split) and n_{th} trees (number of trees to grow).

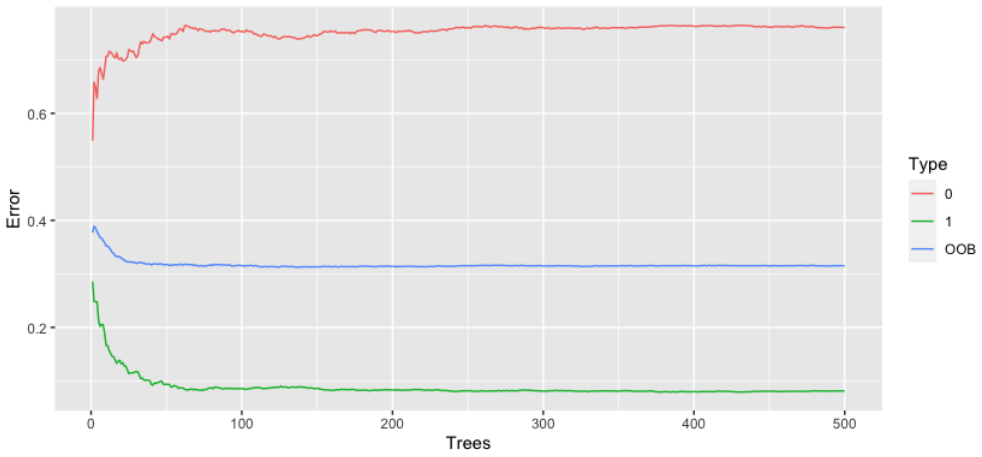

Figure 23: n_{th} Trees

Figure 23 shows that the OOB error rate initially drops and becomes constant. It shows that after about 300 trees, the model could not improve the error. Hence, we use 300 trees to train the model.
The number of variables tried at each split, m_{try} [1] in this result is 4.

$$m_{try} = \sqrt{p} = \sqrt{23} = 4.795 \tag{17}$$

where p is the number of the features. The training set has 23 features, indicating that the model is based on 4 m_{try} ($\sqrt{23} \approx 4$ or 5).

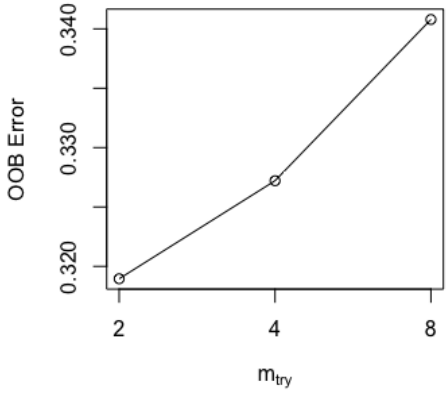

Figure 24: m_{try}

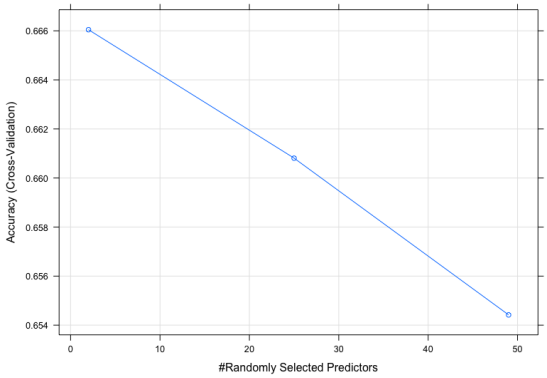

Figure 25: 10-fold cross-validation

Figure 24 clearly shows that the OOB error rate is minimal when m_{try} is 2. The OOB error rate continues to rise, reaching a peak at m_{try} of 8. When $m_{try} = 2$ is selected, the OOB error rate decreases from 33.07% to 31.78%.
We also apply the 10-fold cross-validation method to evaluate the model performance (Figure 25). Similarly, m_{try} of 2 has the highest accuracy.

According to Figure 26, Mean Decrease Accuracy contains a measure of the extent to which the variable improves the accuracy of the forest in predicting the classification [8]. Higher values indicate that the corresponding variable improves the prediction.

Mean Decrease Gini measure of variable importance based on the Gini impurity index (14) [7]. The features are in agreement with the results of the permutation-based variable importance and the Gini importance. The higher the value of the mean decrease, the higher the significant of the variables in the model.

Also, x_{13} (PLU20D) is the most important variable when classifying the data. By looking at the feature importance, we decide that dropping x_6 (GASPLU60D) and x_{16} (WINREP60D) is possible as they do not contribute enough to the prediction process.

Figure 27 shows the model after applying 10-fold cross-validation. Similarly, it shows that x_{13} is the most important variable where SubtypePREFAB is the least important variable.

Figure 28 and Figure 29 shows the marginal effect of x_{13} (PLU20D) on class probability. When $x_{13} < 2$ it tends to predict class 0 more strongly than when $x_{13} > 2$.

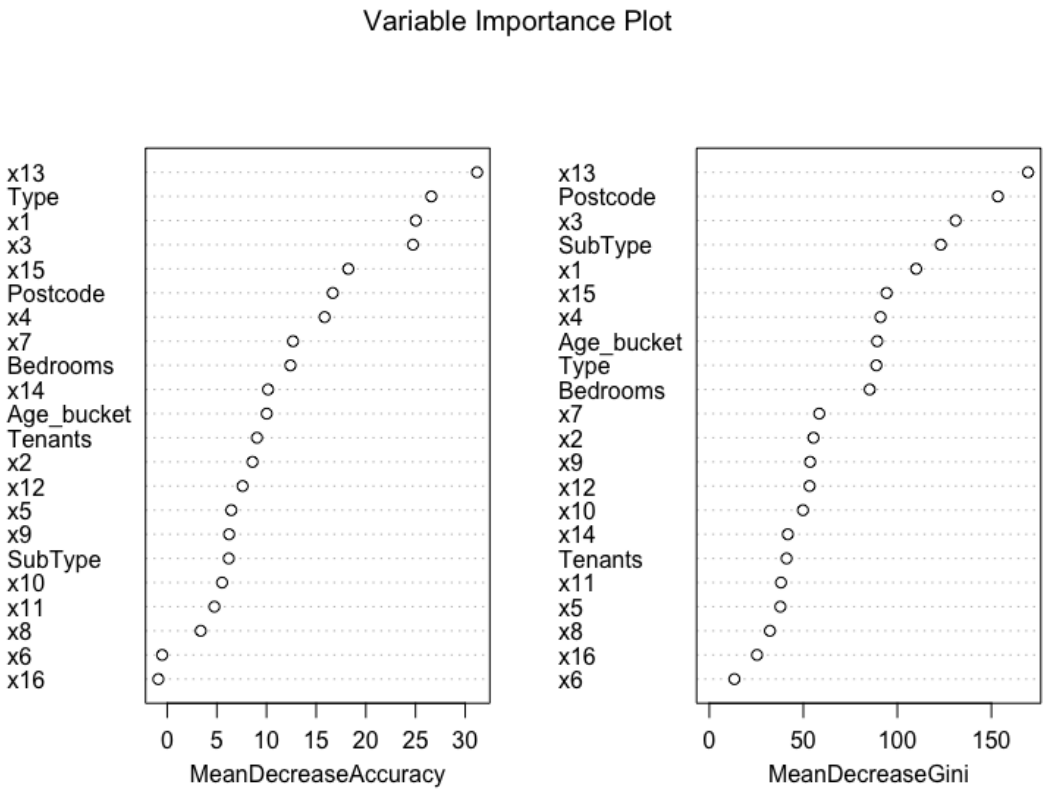

Figure 26: Variable Importance

Predicted Accuracy

Figure 30 plots the distribution probability of the prediction class, which is used to determine whether the model is able to distinguish correctly classified and wrongly classified samples according to the quality of classification. In the case of value 0, there are not only correctly classified samples but also wrongly classified samples. At the same time, there are also correctly classified and wrongly classified samples in the case of value 1.

Figure 31 plots the distribution of votes for each sample in each class. In the case of value 0, the random forest model wrongly classifies most samples with the value of 0 into the category with the value of 1. But in the case of value 1, the model correctly classifies most cases.

Based on Table 14, this model correctly predicted that 1,713 properties will have an emergency repair, and that 237 properties will not have an emergency repair. Thus, the accuracy from the random forest is $(1,713+237)/(1,713+237+734+134)\times 100 = 69.2\%$.

Table 14: Confusion matrix based on the training set of random forest.

Predicted/Actual classes	0	1
0	TN=237	FN=734
1	FP=134	TP=1713

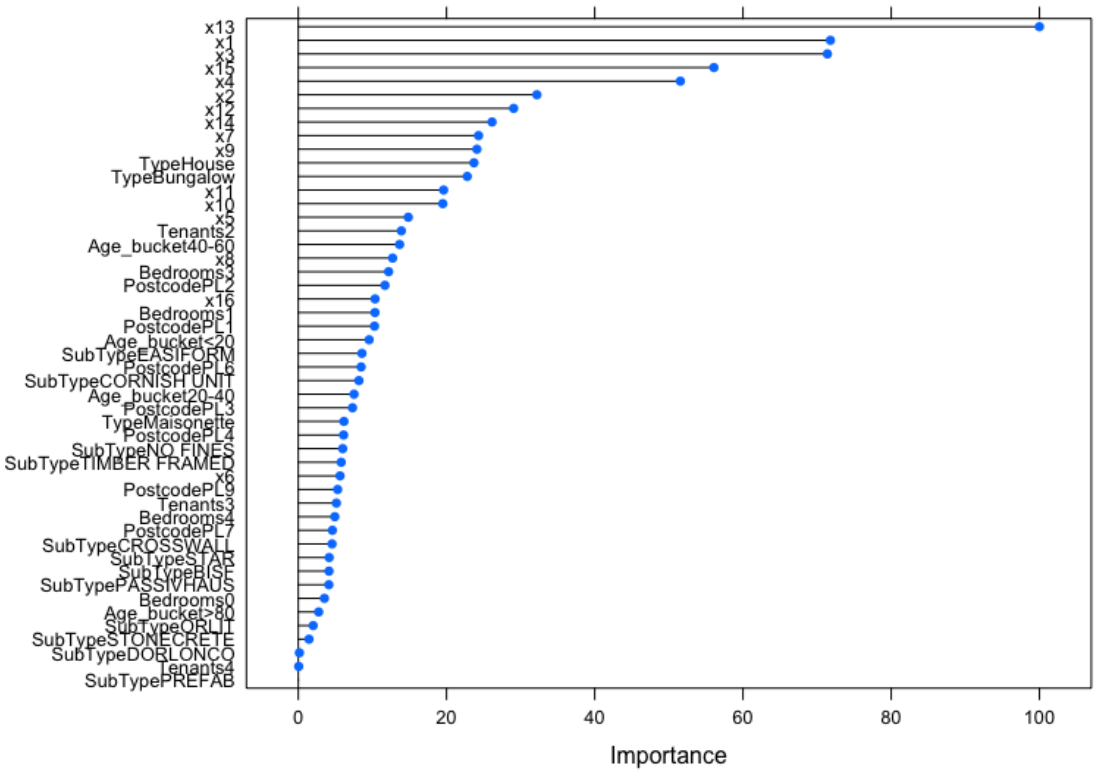

Figure 27: Variable Importance

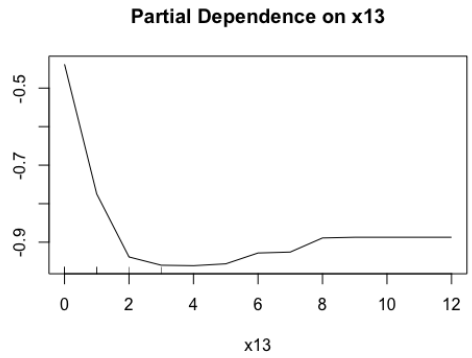

Figure 28: Partial Dependence on x_{13} Class 0.

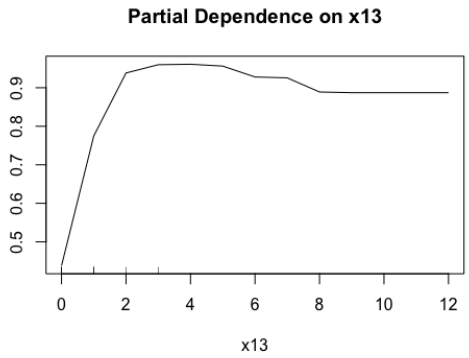

Figure 29: Partial Dependence on x_{13} Class 1.

By performing 10-fold cross-validation, an accuracy of 66.9% is obtained for the random forest model. We see that the accuracy of the random forest model is the highest, and the lowest for the decision tree model among the five models discussed.

Discussion

Logistic regression model does a slightly better job than LDA. LDA assumes that each independent variable is normally distributed but this may not be the case. There are

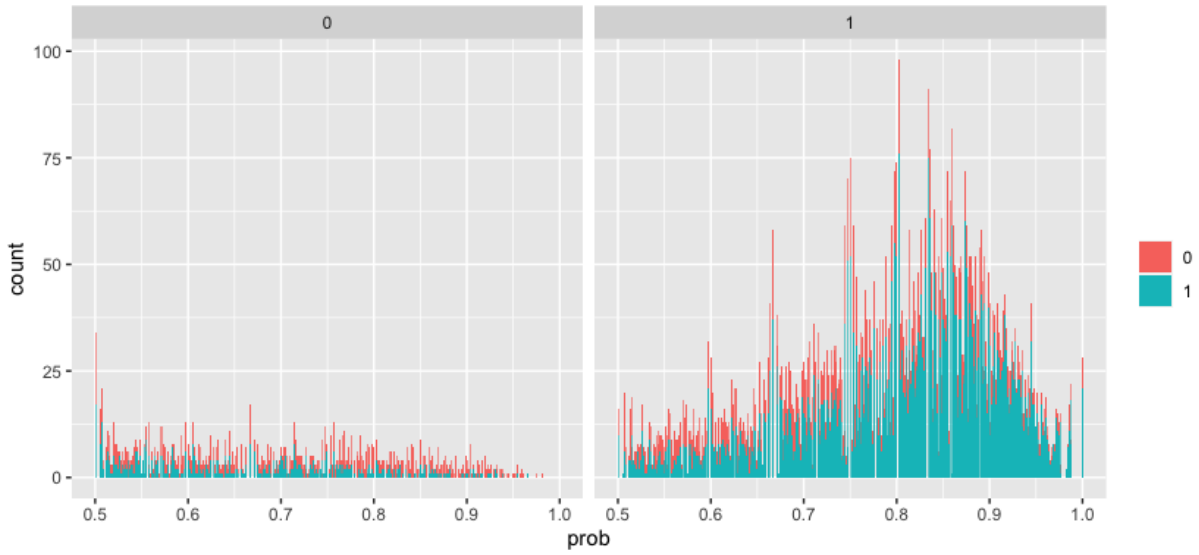

Figure 30: Plot histogram of assignment probabilities to predicted class

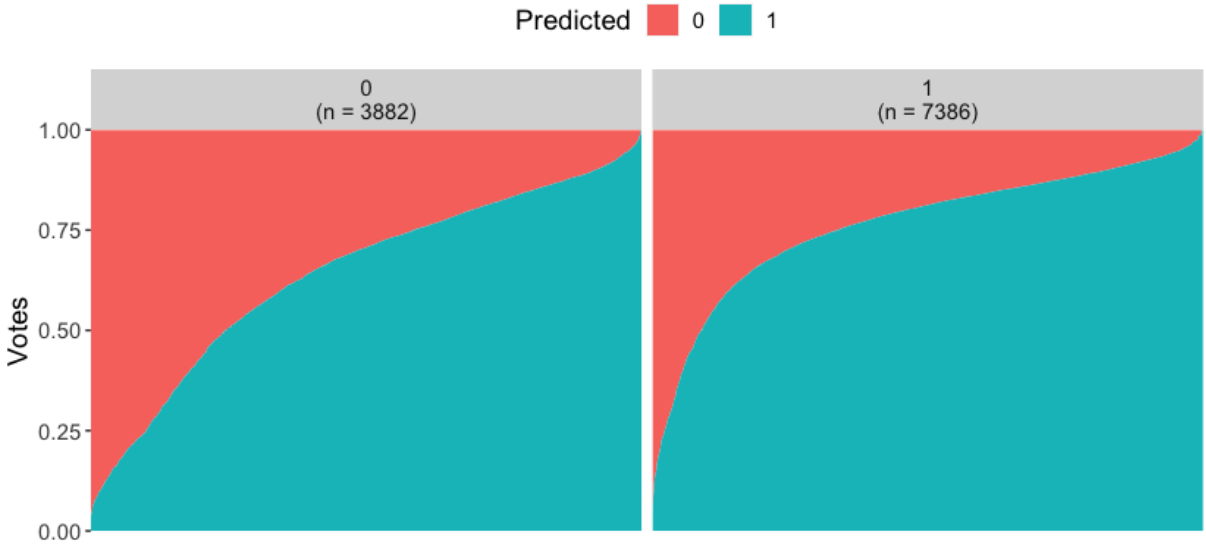

Figure 31: Votes for Random forest

categorical variables that go against this assumption as LDA also assumes that they are continuous. This causes the linear discriminant model to be less stable than the logistic regression model. Furthermore, LDA is more sensitive to outliers. Logistic regression does not need to meet any requirements, making it a flexible and robust model to deal with. If all requirements are met for LDA, then there is a chance that it may classify better. Also, LDA is popular when we have more than two response classes [1].

The *k*-NN algorithm can be used to roughly classify, and there should be no missing values in processing the model. Based on this point, the sample of *k*-NN is different from other models we used for this investigation. The data should be normalised and the calculation of this algorithm needs less time, which are the advantages of this method. In order to have detailed classification and the ranking of the importance of

variables, random forest is used for further model fitting.

The disadvantage of the classification tree is that it is easy to over-fit, which leads to inaccurate prediction results. We want to minimize the error due to data bias and the error due to the variance. Random forests limit over-fitting and errors due to bias by combining classification trees and aggregating those results into a final result [14]. According to the results, the accuracy of the classification tree is 67% and the random forest is 69.2%. This shows that the random forest is better than the single classification tree to grasp the features of the data from the whole and get higher accuracy.

Conclusion

All the models obtained approximately equal accuracy on the predicted probabilities. This may be due to the quality of the dataset. The accuracy could be improved if we had more data to work with. But, higher accuracy does not always indicate a better performance. Sometimes, the improvement in model's accuracy can be due to over-fitting too [15]. Therefore, the models discussed may have performed decently but with its own advantages and disadvantages.

Acknowledgements

We would like to acknowledge Dr Yinghui for her guidance on the project and to Plymouth Community Homes (PCH) with the access to the anonymised data for this study.

References

[1] J. Gareth, D. Witten, T. Hastie and R. Tibshirani, *An introduction to Statistical Learning with Applications in R*, Springer, New York, 2013.

[2] Harshitha Mekala, *Dealing with Missing Data using R*, https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17

[3] Christina, *An Introduction to Logistic Regression for Categorical Data Analysis*, https://towardsdatascience.com/an-introduction-to-logistic-regression-for-categorical-data-analysis-7cabc551546c

[4] *Simple guide to confusion matrix terminology*, https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

[5] https://jairiidriss.medium.com/a-simple-explanation-of-entropy-and-information-gain-decision-tree-classification-machine-99d89d094443

[6] Jason Brownlee, *Linear Discriminant Analysis for Machine Learning*, https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/

[7] *The Forest Model*, http://blog.keyrus.co.uk/alteryxs_r_random_forest_output_explained.html

[8] Palczewska, A., Palczewski, J., Robinson, R. and Neagu, D., 2013. *Interpreting random forest models using a feature contribution method*. 2013 IEEE 14th International Conference on Information Reuse Integration (IRI),.

[9] Cole Eagland, *ISLR8 - Decision Trees in R (Classification)*,https://rstudio-pubs-static.s3.amazonaws.com/442284_82321e66af4e49d58adcd897e00bf495.html

[10] *Building a classification tree in R*, https://davetang.org/muse/2013/03/12/building-a-classification-tree-in-r/

[11] ManjuLa S, *Decision Trees- What, How and Why?*, https://manjula2020.medium.com/decision-trees-what-how-and-why-e6e9577c71f6

[12] Cory Maklin, *Random Forest In R*, https://towardsdatascience.com/random-forest-in-r-f66adf80ec9

[13] Dima Shulga, *5 Reasons why you should use Cross-Validation in your Data Science Projects*, https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79

[14] Neil Liberman, *Decision Trees and Random Forests*, https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991

[15] Sunil Ray, *8 Proven Ways for improving the "Accuracy" of a Machine Learning Model*, https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/

[16] *Explanation of the Decision Tree Model*, https://webfocusinfocenter.informationbuilders.com/wfappent/TLs/TL_rstat/source/DecisionTree47.htm