Scientific Research Publishing

Open Journal of Statistics, 2020, 10, 600-610
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

# Analysis of the Resolution of Crime Using Predictive Modeling

**Keshab R. Dahal[1*], Jiba N. Dahal[2], Kenneth R. Goward[3], Oluremi Abayami[4]**

[1]Department of Statistics, Truman State University, Kirksville, MO, USA
[2]Department of Physics, Truman State University, Kirksville, MO, USA
[3]Department of Mathematics, SUNY Cortland, Cortland, NY, USA
[4]Department of Mathematics, Northwood University, Midland, MI, USA
Email: *kdahal@truman.edu

**How to cite this paper:** Dahal, K.R., Dahal, J.N., Goward, K.R. and Abayami, O. (2020) Analysis of the Resolution of Crime Using Predictive Modeling. *Open Journal of Statistics*, **10**, 600-610.
https://doi.org/10.4236/ojs.2020.103036

**Received:** June 7, 2020
**Accepted:** June 27, 2020
**Published:** June 30, 2020



## Abstract

There has been evidence of crime in the US since colonization. In this article, we analyze the crime statistics of San Francisco and its resolution of crime recorded from January to September of the year 2018. We define resolution of crime as a target variable and study its relationship with other variables. We make several classification models to predict resolution of crime using several data mining techniques and suggest the best model for predicting resolution.

## Keywords

Machine Learning, Classification Model Comparison, Predictive Modeling, Resolution of Crime

## 1. Introduction

On a daily basis, all manners of residents in the United States are affected by crimes. Crime rates vary over time, reaching its peak between the 1970s and early 1980s. According to the FBI [1], there are two types of crimes in the USA namely violent crime and property crime. Crimes such as murder, manslaughter, and rape are described as violent crime whereas crimes such as burglary, larceny, and vehicle theft belong to property crime.

In order to implement law and order effectively, one must analyze the crime statistics and should minimize the number of unsolved crimes as low as possible. In this article, we analyze the crime statistics of San Francisco and its resolution (resolved or not resolved) of crime recorded from January to September of the year 2018. We define resolution of crime as a target variable and study its relationship with other variables. We make several predictive models to predict

"Resolution of crime" using several machine learning techniques and suggest the best model (or models).

Several authors have defined machine learning in their own way. One of the common ways to define machine learning is: Technology uses for the development of computer algorithm with the ability of imitating the intellectuality of human beings is known as machine learning. It is produced from the ideas of the different fields such as Computer Science, Information Theory, Statistics and Probability, Artificial Intelligence, Psychology, Control Theory and Philosophy [2] [3] [4].

It has been a very challenging question which model type to apply to a machine learning task in order to make a precise prediction. Every model has some merits and demerits [5]. It can be difficult to compare the relative merits of the models. In this paper, five different supervised classification machine learnings: Logistic Regression (LR), Classification Tree (CART), Linear Discriminant Analysis (LDA), Quadrilateral Discriminant Analysis (QDA), and K-Nearest Neighbor (KNN) are implemented. We use these five classification models to predict the resolution of crime. Finally, the performance of the algorithms is compared to select the best model.

In section 2, we discuss data description and preprocessing. Different classification machine learning will be discussed in section 3. In section 4, we compare models and select the best model based on their performance. In section 5, we summarize the main findings and conclude the journal.

## 2. Data Description and Preprocessing

### 2.1. Data Source

In this study, we use the publicly available dataset that we obtained from San Francisco Police Department Incident Reports from January to September of the year 2018, which has information of 111,531 official crimes. This project started on October 2018; therefore, the only data available was from January to September of 2018. Every entry in the dataset contains information about a crime. The dataset contains 26 variables and 111,531 observations. The detail information of the dataset with variable name, type, and level are available in [6].

### 2.2. Data Cleaning

In the case of a large dataset, learning the dataset is not useful unless the unwanted features are removed since an irrelevant and redundant feature does not add anything positive and new to the target concept [7]. Before implementing machine learning algorithms to our dataset, we went through a series of prepossessing steps.

- Dropping irrelevant features

The feature which has almost negligible effect on the response variable is called irrelevant feature. One of the common examples of irrelevant feature is serial number. In data mining, there are many features of selection methods

such as "Filter Method", which automatically drop the irrelevant features. In general, we use the feature selection method if you have a huge number of features in hand. However, since our dataset has only 26 features, it is not difficult to identify the irrelevant features and omit them from the further process. The variables: Incident Code, Incident Number, Incident ID, Row ID, Report Type Code, and CAD (Computer Aided Dispatch) Number are irrelevant identifiers, so they are omitted.

- Dropping redundant features

The variable Datetime is rejected since it gives the same information as Incident Day of the week and Incident Time. Report Datetime, Report Type Code, and Report Type Description are rejected since we care when the crime was committed, not reported. Point provides the same information as Latitude and Longitude, so it is rejected. The variables Analysis Neighborhood and Police District give the same information. The Analysis Neighborhood has missing value as opposed to Police District so we keep Police District and reject Analysis Neighborhood. The variables Incident category, Incident Subcategory, Incident Description give the same information, so we keep the variable Incident category as an input variable and the other two are rejected.

- Imputing missing values

Missing data is a common problem in data mining. Rates of less than 1% missing data are generally considered trivial, 1% - 5% are manageable. However, 5% - 10% requires sophisticated method to handle, and more than 15% may severely impact any kind of interpretation [8]. The variables CNN (The unique identifier of the intersection for reference back to other related basemap datasets), Latitude, Longitude, and Supervisor District have 5575 missing values. Approximately 5% of the data are missing in our datasets so it is not reasonable to ignore missing data and delete from dataset. Several methods for imputation of missing data together with their merits and demerits have discussed [9]. Missing values of our datasets include both numeric and categorical so the reliable way to impute is K-nearest neighbors (KNN). KNN algorithm is the algorithm most useful for any kind of missing data because it takes missing data within its closet k neighbors in the multi-dimensional space. We imputed the missing values using KNN method explained in [10] with k = 10.

- Data transformation

The variable Filed Online is either TRUE or blank in the original data, so it is converted to TRUE/FALSE to represent whether a report was filed online or not.

The variable Incident category is a characteristic variable with 39 subcategories which is not feasible to interpret. We realized that more meaningful approach is to collapse the categories into fewer, large groups: Assault, Burglary, Larceny theft, Non-criminal, and Others.

We have used the case when command in dplyr package of R to change the level of variables Filed Online and Incident category.

The variable incident date was a categorical variable with standard US date format (MM/DD/YYY), which gives the information of the incident starting

from 1st January to 24th September. In order to make the analysis fruitful and feasible, we have extracted the incident month from the incident date and converted the incident date to incident month with 9 different categories from January to September using case when command explained above. Similarly, Incident Time was a categorical variable with time format HH: MM. This is decomposed into four categories: Morning, Afternoon, Evening, and Overnight. We decomposed such that: 6 am-noon as Morning, noon-6 pm as Afternoon, 6 pm - 10 pm as Evening, and midnight - 6 am overnight.

The variable Resolution is a categorical variable with 6 classes: Open or Active, Cite or Arrest Adult, Cite or Arrest Juvenile, Exceptional adult, Exceptional Juvenile, and Unfounded. We define classes; Cite or Arrest Adult, Cite or Arrest Juvenile, Exceptional adult, Exceptional Juvenile as Resolved and other two classes; Open or Active, and Unfounded as Unresolved so that the variable Resolution become binary with 1 for resolved and 0 for unresolved. We decided to take this as a Target variable. The brief summary of the cleaned data with role, type, and level is summarized in Table 1.

- Encoding Categorical Feature

Feature engineering is a crucial part of machine learning. Since the implemented algorithm is only able to read numerical values, it is extremely important to encode that the categorical features are transformed into numerical values. Many statistical learning algorithms such as LDA, and QDA require as input a numerical feature matrix. When categorical variables are present in the data, feature engineering is needed to encode the different categories into a suitable feature vector [11]. We have transferred the categorical variables: Incident Month, Incident Time, Incident Day of Week, Incident Category, and Police District to numerical variables by simply replacing categories by counting numbers.

**Table 1.** Summary of the variables name, role, type and level of cleaned data.

| Variable Name | Variable Role | Variable Type | Variable Level |
|---|---|---|---|
| CNN | Input | Numeric | Interval |
| Latitude | Input | Numeric | Interval |
| Longitude | Input | Numeric | Interval |
| Incident Month | Input | Characteristic | Nominal |
| Incident Time | Input | Characteristic | Nominal |
| Incident Day of Week | Input | Characteristic | Nominal |
| Incident Category | Input | Characteristic | Nominal |
| Police District | Input | Characteristic | Nominal |
| Supervisor District | Input | Numeric | Interval |
| Filed Online | Input | Characteristic | Binary |
| Resolution | Target | Characteristic | Binary |

- Feature Scaling

Since most machine learning algorithms for example KNN, use Euclidean distance between two data points; data sets containing various ranges are a problem. Features need to be accurate. Due to this, feature scaling is utilized to repress the explained effect to gather all of the features into the same magnitude [12].

To scale the features of the dataset, standardization has used. The formula used to calculate the standardization is as follows:

$$z = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where $z$, min $(x)$, and max $(x)$ are standardized input, minimum, and maximum values for the features, respectively.

### 2.3. Data Partition

In this part of the preprocessing stage, the data is split into two parts: training and testing data in the ratio 3:1. We have used the sample command of R to select 75% of the entire dataset. This random sample is taken as train data. The remaining 25% of the data is considered as test data. The main purpose of the splitting data is to avoid overfitting. There might be the case where the machine learning algorithm performs exceptionally well in the training dataset, however, performs badly in the testing dataset.

## 3. Machine Learning Algorithms

There are various machine learning algorithms available to solve the classification problems such as Logistic Regression, Neural Network, and Support Vector Machine. However, our research is limited to the following machine learning algorithms.

### 3.1. Logistic Regression

Logistic Regression (LR) Model is used for predicting binary outcomes. It is a statistical model that in its basic form uses as a sigmoid function to model a binary response variable, taking on values 1 and 0 with probability $\pi$ and $1 - \pi$ respectively. A logistic regression model is given below as:

$$\text{logit}\left(\Pr\left(Y=1\right)\right) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{2}$$

where,

$$\text{logit}\left(\Pr\left(Y=1\right)\right) = \ln\left(\frac{\Pr\left(Y=1\right)}{1-\Pr\left(Y=1\right)}\right) \tag{3}$$

LR is one of the most popular and common method that has been used for a long time to solve classification problem especially when the response variable is binary. Due to simplicity and convenience, the first method that comes in the mind of most statistical is LR. We have fitted the logistic regression model using

the glm commands of R package as explained in [13].

### 3.2. Linear Discriminant Analysis

Fisher Linear Discriminant Analysis (also called Linear Discriminant Analysis (LDA)) is a method used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification [14].

Though their motivation differs, the logistic regression and Linear Discriminant Analysis (LDA) are closely connected. The only difference between these two models is the way their parameters are estimated. In Logistic Regression, the parameters are estimated using maximum likelihood, whereas in LDA method, the parameters are computed using the estimated mean and variance from the normal distribution. In LDA method, we assume that the variables follow Gaussian distribution with common covariance matrix. If this assumption is met, LDA outperforms Logistic Regression. Conversely, Logistic Regression outperforms LDA if these assumptions are not met. We fit the LDA model using R command lda of the MASS package similar to the procedure explained in [5].

### 3.3. Quadrilateral Discriminant Analysis

Quadrilateral Discriminant Analysis (QDA) is a supervised machine learning in which a quadratic decision boundary classifier is used to differentiate the class. QDA serves as a compromise between LDA and Logistic Regression approach and the nonparametric KNN method. QDA is more flexible than LDA and Logistic Regression as its decision boundary is quadratic but less flexible than KNN. A QDA model is fitted using R command qda of the MASS packages like the procedure explained in [5].

### 3.4. Classification Tree

Classification trees are a powerful alternative to more traditional approaches of land cover classification. Trees provide a hierarchical and nonlinear classification method and are suited to handling non-parametric training data as well as categorical or missing data. By revealing the predictive hierarchical structure of the independent variables, the tree allows for great flexibility in data analysis and interpretation [15]. Classification tree is simple and useful for interpretation. It is a statistical model which is used to predict a qualitative response. In this model, we predict that each observation belongs to the most commonly occurring class of training observations in the region which it belongs to. A Classification tree with the best value of complexity parameter is fitted using R package rpart similar to the procedure explained in [16].

### 3.5. K-Nearest Neighborhood

KNN model takes a completely different approach than the other classification

models. To fit KNN model, no assumption is needed. In fact, it is completely nonparametric. KNN can outperform other classification models if the assumptions are not met. We fit the KNN model using R packages Class similar to the procedure explained in [10].

## 4. Model Comparisons

To determine which model has the better performance, they were trained on the training dataset and fit to the test dataset to retrieve the following matrices: Sensitivity, Specificity, and Accuracy. We compute the confusion matrix for each model as shown in **Table 2**.

The proportion of the actual resolved case that is correctly predicted as resolved is called sensitivity. It is also called true positive rate (TPR) and is given in Equation (4).

$$\begin{aligned}\text{Sensitivity} &= \text{True positive rate}(\text{TPR}) \\ &= \frac{\text{True positive}(\text{TP})}{\text{True positive}(\text{TP}) + \text{False negative}(\text{FN})}\end{aligned} \tag{4}$$

The proportion of the actual unresolved case that is correctly predicted as unresolved is called specificity. It is also called false positive rate (FPR) and is given in Equation (5).

$$\begin{aligned}\text{Specificity} &= \text{False positive rate}(\text{FPR}) \\ &= \frac{\text{True negative}(\text{TN})}{\text{True negative}(\text{TN}) + \text{False positive}(\text{FP})}\end{aligned} \tag{5}$$

The proportion of the cases that is predicted accurately is called the accuracy and is defined by Equation (6).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{6}$$

The model with higher statistics: sensitivity, specificity, and Accuracy is considered as a better model. **Table 3** summarizes such statistics. The sensitivity of

**Table 2.** Confusion matrix.

| | Actual Resolved | Actual Unresolved |
|---|---|---|
| Predicted Resolved | TP | FP |
| Predicted Unresolved | FN | TN |

**Table 3.** Model comparison of five models.

| Model method | Sensitivity | Specificity | Accuracy |
|---|---|---|---|
| Logistic Regression | 0.1712 | 0.9585 | 0.7685 |
| Classification tree | 0.6119 | 0.8112 | 0.7864 |
| LDA | 0.003715 | 0.9974 | 0.7576 |
| QDA | 0.1851 | 0.9476 | 0.7635 |
| KNN | 0.4187 | 0.8819 | 0.7701 |

models: LR, LDA, and QDA are less than 18%, which is very low so they can't be considered as a better model because less than 18% of the time, they correctly predict the actual resolved cases to be resolved cases. On the flipside, sensitivity of Classification tree is 0.6119 which is highest among the models.

Specificity of all models are reasonable. All models were able to attain at least 88%. The accuracy of the Classification tree is 0.7864, which is the highest. So the Classification tree is considered as a better model.

## 5. Results

We compared different classification machine learning algorithms for predicting the resolution of crime using the publicly available dataset that we obtained from San Francisco Police Department Incident Reports from January to September of the year 2018. The Classification tree followed by Logistic Regression outperforms the other three models: Liner Discriminant Analysis, Quadrilateral Discriminant Analysis, K nearest neighborhood.

A possible cause is that KNN suffers from the poor performance whenever the class distribution of the Resolution is skewed [17]. Most of the voting will raise conflict when there are huge class that dominates prediction. There will also be a tendency for new data to be voted into additional popular classes. **Figure 1** verifies the fact that the number of unsolved cases is almost four and half times more than the number of solved cases. As a result, it is unsuitable to use KNN in this dataset.

It is worth noting that in models: Liner Discriminant Analysis and Quadrilateral Discriminant Analysis, the sensitivity is very low, less than 20%. This is likely due to the fact that the dataset failed to meet Gaussian requirement. It can be seen from **Figures 2-4**, several variables fail to follow Gaussian distribution. The feature Longitude is skewed to the left as shown in **Figure 2**. Similarly, the variables Latitude and CNN are skewed left and skewed right with possible

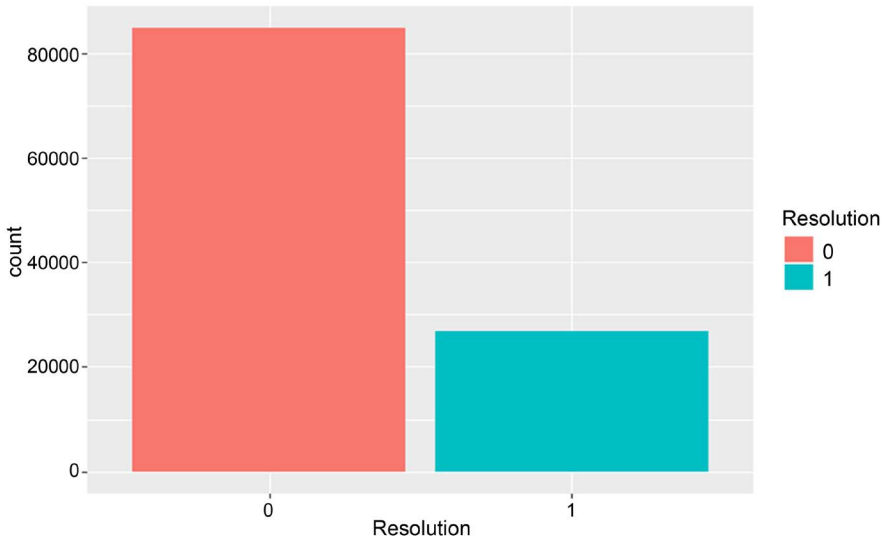


**Figure 1.** Distribution of resolution.

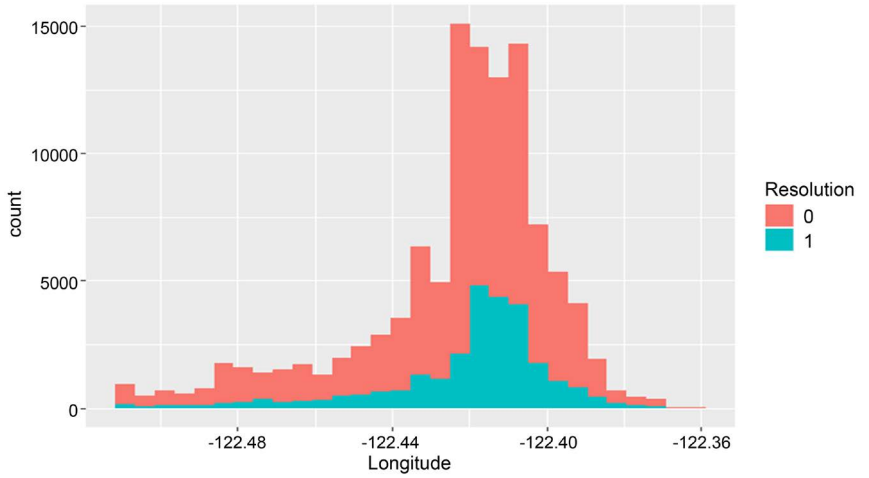


**Figure 2.** Histogram of longitude.

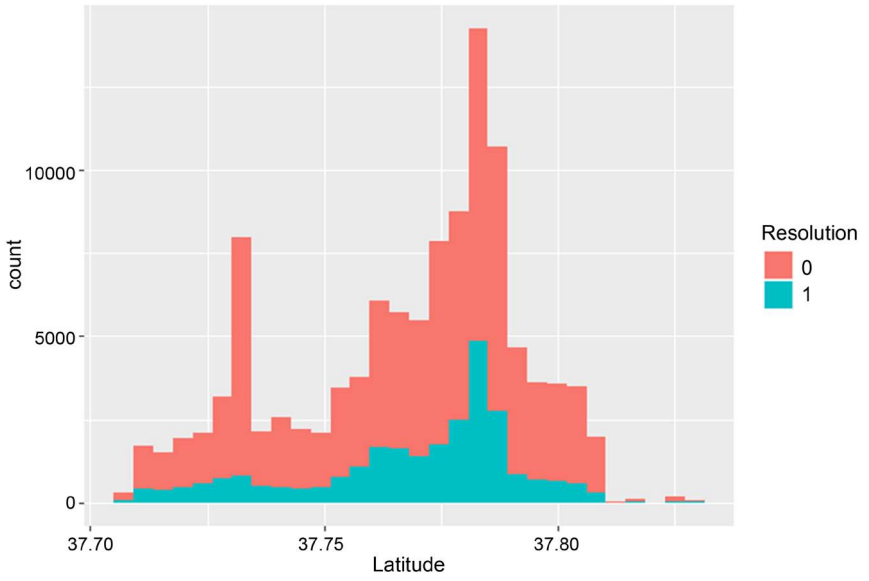


**Figure 3.** Histogram of latitude.

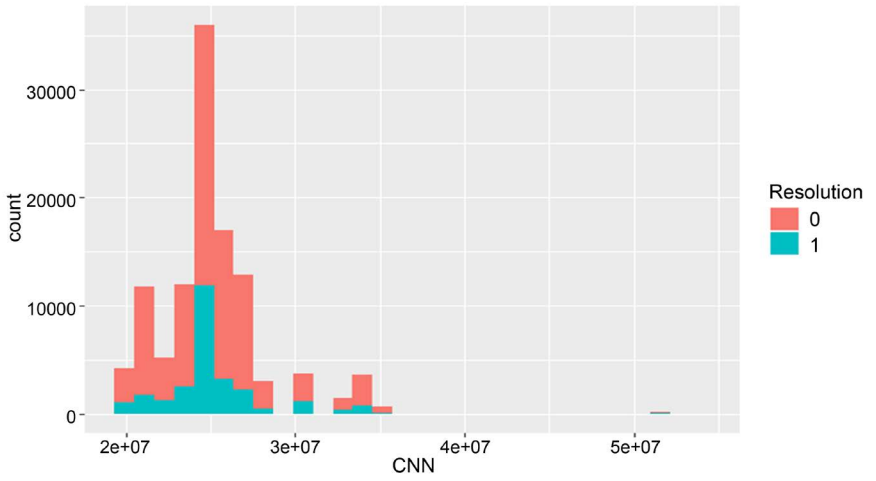


**Figure 4.** Histogram of CNN.

outlier as shown in Figure 3 and Figure 4 respectively. Another possible reason for the poor performance is the categorical features transferred into counting numbers.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] 2017 Crime in the United States. https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/topic-pages/property-crime

[2] Mitchell, T.M. (1997) Machine Learning. McGraw-Hill Higher Education, New York.

[3] Alpaydin, E. (2020) Introduction to Machine Learning. MIT Press, Cambridge.

[4] Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer, Berlin.

[5] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning. Vol. 112, Springer, New York, 3-7. https://doi.org/10.1007/978-1-4614-7138-7

[6] Police Department Incident Report of City and County of San Francisco. https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783

[7] Guyon, I. and Elisseeff, A. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**, 1157-1182.

[8] Acuna, E. and Rodriguez, C. (2004) The Treatment of Missing Values and Its Effect on Classifier Accuracy. In: *Classification, Clustering, and Data Mining Applications*, Springer, Berlin, Heidelberg, 639-647. https://doi.org/10.1007/978-3-642-17103-1_60

[9] Van Buuren, S. (2018) Flexible Imputation of Missing Data. CRC Press, Boca Raton. https://doi.org/10.1201/9780429492259

[10] Crookston, N.L. and Finley, A.O. (2008) yaImpute: An R Package for kNN Imputation. *Journal of Statistical Software*, **23**, 16 p. https://doi.org/10.18637/jss.v023.i10

[11] Cerda, P., Varoquaux, G. and Kégl, B. (2018) Similarity Encoding for Learning with Dirty Categorical Variables. *Machine Learning*, **107**, 1477-1494. https://doi.org/10.1007/s10994-018-5724-2

[12] Asaithambi, S. and Why, H. (2017) Why, How and When to Scale Your Features. https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e

[13] Manning, C. (2007) Logistic Regression (with R) Changes.

[14] Li, C. and Wang, B. (2014) Fisher Linear Discriminant Analysis. CCIS Northeastern University.

[15] Hansen, M., Dubayah, R. and DeFries, R. (1996) Classification Trees: An Alternative to Traditional Land Cover Classifiers. *International Journal of Remote Sensing*, **17**, 1075-1081. https://doi.org/10.1080/01431169608949069

[16] Therneau, T., Atkinson, B., Ripley, B. and Ripley, M.B. (2015) Package "rpart". http://cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf

[17] Coomans, D. and Massart, D.L. (1982) Alternative k-Nearest Neighbour Rules in Supervised Pattern Recognition: Part 1. k-Nearest Neighbour Classification by Using Alternative Voting Rules. *Analytica Chimica Acta*, **136**, 15-27. https://doi.org/10.1016/S0003-2670(01)95359-0