Predicting of Hosting Animal Centre Outcome Based on Supervised Machine Learning Models

MSc Research Project
MSc Data Analytics

Sushant Parte
Student ID: X18137440

School of Computing
National College of Ireland

Supervisor: Dr. Vladimir Milosavljevic

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sushant Prabhat Parte		
Student ID:	18137440		
Programme:	MSc Data Analytics	**Year:**	2019-2020
Module:	MSc Research Project		
Supervisor:	Dr. Vladimir Milosavljevic		
Submission Due Date:	12/12/2019		
Project Title:	Prediction of Hosting Animal centre Outcome based on Supervised Machine Learning models		
Word Count:	**7316**	**Page Count: 22**	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ………

Date: ………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Hosting Animal Centre Outcome Based on Supervised Machine Learning Models

Sushant Parte
X18137440

Abstract

[1]In 2010 Austin passed a plan known as No kill implementation plan which is maintained by Austin Animal center and pursued a 90% live outcome goal. Today Austin is one of the largest no kill animal county in world. These research project helps by the means of data science to develop an approach to its best capabilities to increases the live rate by different means in future. In this research article we used four different supervised learning classification models with feature engineering and implementing the vectorization process. The dataset was cleaned, and monitoring data was created to implement the models. The models implemented were logistic regression, Neural Network, XGboost and Random forest with K-fold holdout cross validation to calculate the and predict the outcome type of animals from Austin animal center outcome dataset. The models evaluated with different evaluation metrics like accuracy, logarithmic loss, sensitivity and specificity providing the output as XGboost outperformed compared to all the other classification models with accuracy of 65.33%. the prediction and actual figures were determined by building confusion metrics.

Keywords- *Classification, XGboost, Neural Network, Logistic Regression, Random Forest, Animal center.*

1. Introduction

A pet is a companion animal kept primarily in company of a person for entertainment or certain act of compassion. People get pets for protection, emotional care, physical attractiveness, relive from stress and exertion and love. In recent times, we are surrounded by most of the two popular pets as cats and dogs, which are popular pets among all animals. These two animals provide human beings the support and compassion mentally and emotionally. Having cat or dog as pet keeps the human fit as every dog needs daily walks in order to stay fit and healthy, remove the feeling of loneliness and the best part is stopping the growing children to get into any allergies as asthma[2]. The research project contains the outcomes to be decided how the pets are being sheltered and the behavioral pattern for the same in the city of Austin. Austin is a capital city in of the U.S. state of Texas. Austin has an no kill animal shelter which provides shelter for the stray animals or lost animals. It is an animal shelter that does not allow people to kill animals and provide treatment for animals. The shelter provides 16,000 and more animals shelter on annual basis. The main objective of

[1] http://www.austintexas.gov/blog/no-kill-austin
[2] https://bluebuffalo.com/articles/pets/health-benefits-of-furry-family-members/

the center is to provide permanent shelter for the adoptable animals through adoption. The animal center provides remedies like food, water and vaccination. The research project uses the Austin animal shelter outcome data to classify and predict different outcomes obtained for animals using machine learning models.

The classification algorithms used for the supervised learning approach are logistic regression, Neural Network, Random Forest and XGboost. The models performed well in classifying the results and prediction of the classification were evaluated using confusion matrix. Whereas the model performance was tested using the evaluation metrics like accuracy, logarithmic loss, sensitivity and specificity.

1.1 Motivation

The research provides implementation of classification models to increase the rate of adoption. Increase in the adoption rate help the residents in improvement in social and economic environment. Adoption of pets will create a pet friendly environment and provide an ability for the city to attract new residents[3]. Addition to tourist attraction it will raise the cost paid to veterinaries and pet care services. The research also supports the factor to be considered for other centers to increase the rate of adoption.

1.2 Research question

"Can supervised machine learning classifiers outperform the expectation of Austin Animal shelter to determine which animals are probably to be euthanized and additionally find trends in features to increase the chances for adoption?"

The basic objective of the research question is to correctly classify the outcome of animals and reduce the measure of euthanized of animals using classification models.

1.3 Research objective

Objective 1- To improve the rate of adoption at Austin animal Center.

Objective 2- To safeguard and help animals using power of data science.

Objective 3- Provided increase rate in adoption raises the mental health and stability of residents.

2. Related work based on Outcomes in different fields

For performing the research, we will study different classification and various feature engineering methods implemented to develop an approach to perform the research in better way. In this section will relate the different work performed by different authors and researchers to develop an evidence to support the performed work to its best.

[3] https://www.davisenterprise.com/forum/opinion-columns/the-economics-of-a-no-kill-animal-shelter/

2.1 Study of Prediction of different outcomes

Predicting the outcomes with different features and providing accurate information is a challenging task which could be reviewed by [1]. The outcome predicted in the research paper has highlighted cancer as the major disease to be categorized with different subtypes to upgrade the necessity in cancer research with the help of machine learning (ML) methods. The research mainly focuses on dimensionality reduction and feature classification to perform task like classification and prediction. The models implemented were decision tree, artificial neural network and support vector machine. The research performed on huge size of dataset, but best informative features were considered. The research proved that ANN and SVM proved to be the best model for classification prediction. The research which is related to biological environment for gene selection and prediction using logistic regression. The dataset imported reflects some relevant and irrelevant features in the research which were filtered using Gibbs sampling method to discover important genes. The predicted genes showing high probability of having cancer was done by logistic regression with high accuracy which can be reviewed form the article [2]. The research [3] performed, predicted the best model to be used for classification and prediction by comparing the results of logistic regression and neural network with different performance and evaluation metrics. (CART) Classification and regression was also implemented to derive the prediction results. ROC curves were plotted to compare the best prediction model which proved that logistic regression and CART were the best among the three.

The concept of classification and prediction can be applied in the field of sports which can be reviewed by article [4]. The research paper reviews about usage of SVM with fuzzy membership functions which enhances and smoothens the decision-making surface. Cross validation method was used to compute the accuracy for both the models and the most accurate model was selected for improving the decision making for coaches and sport committee to select the best player. The prediction of death or alive for human beings can be reviewed with machine learning algorithm by the article [5]. The research work proposes a rule-based model to compare the accuracies of individual models like support vector machine, decision tree and logistic regression. The reason for comparison to predict the accurate result for cardiovascular diseases the performance result was evaluated using specificity, sensitivity and accuracy. The research article [6] proposes 1000 coronary heart disease cases which concluded that SVM showed the best prediction accuracy with 92.1% following neural network and decision tree with 91% and 89.6% respectively. The research proposed a 10-fold cross-validation to provide insight with different data having 11 attributes. The article [7] reviews collection of 102 cases having Coronary heart disease (CHD) where each case is diagnosed by TCM as which syndrome and corresponding nine NEI specifications are measured. This can be done using 4 distinct algorithms as Decision Trees, Support vector machine, neural network and logistic regression. The compared accuracy for all the models concluded that SVM has the best prediction accuracy and can be used to classify and predict the cases suffering from CHD.

The field of education also can be induced in the field of machine learning of prediction which can be reviewed from the article [8]. The article reviews about the growing low graduation rates in U.S. higher education system. The article reviews about the predictive model of students' graduation outcomes based on ensembled machine learning models and evaluating proposed technique using experimental study. The prediction of on time graduation of student was accurately predicted with SVM ensemble model. The error rate can be reduced by 5-fold cross validation for each of the predictive model can be reviewed from the article [9]. The article reviews the comparison of error rate of training error which is downward approach and 5-fold or 10-fold cross validation error rate which is upward bias approach. These two families error rate is being investigated in the reviewed paper. Therefore, the article concludes that cross validation estimated lowest error rate. The importance of K-fold cross validation method used for mainly classification model can be reviewed from the article [10]. The article reviews two methods k-fold method for large dataset and leave-one-out cross validation method for smaller dataset. The article reviewed both the approaches are popular approaches for evaluating the performance of classification algorithm. The success and failure rate of prediction model can be understood by [11] which describes the featuring engineering applied on the dataset to derive proper evaluation rates for prediction of the software projects success or failure in a Japanese software sector. The resultant algorithm used extracted an accuracy of 77.8%. This made easier for IT vendors to make decision in investment of finance project. Neural network provides a better performance in classification using the back propagation algorithm and feature selection reviewed by the article [12]. The article tested the diagnosis of heart disease on certain test cases and for classification reduced the number of attributes from 13 attributes to 8 attributes using information gain. The model aims to classify to presence and absence of heart diseases and is evaluated using the confusion matrix. A comparative study of different algorithms develops in improving the decision making in the usage of certain algorithm in that field reviewed from the article [13] which compares random forest and SVM two data mining algorithm for protein function prediction in field of bioinformatics. The data set sued consisted of different enzymes which were classified found with overall accuracy 88.9% for SVM and 53.9% for random forest. The results were evaluated using accuracy, sensitivity, specificity and precision.

2.2 Study of Feature Engineering for Predictive analysis

Feature engineering for predicting the model appropriately is the trending approach inculcated for large and complex dataset which can be reviewed from the article [14]. The article reviews simplification of dataset having multiple input and multiple output control problems. This approach results in dimensionality reduction and multivariant regression algorithms. The results for this approach reduce the hardware use and implementation cost which is implemented by principal component analysis and dynamic approach of the building model. Therefore, the evaluation of the models is strongly correlated with the performance of the models with better approach. The use of this approach is compared to other related works and traditional approach which in return provides scalability and computational efficiency for proposed models. There are various feature engineering approaches for predicting the models correctly but he best feature which can be used is reviewed by the article [15]. The article describes the best empirical research to be used to demonstrate the different engineering features that are demonstrated to a machine learning model. The research article reveal that different model responds to different engineering feature. The experiment performed reveals to what extend the models are capable of synthesizing needed features. The research paper concludes that ensemble model performs better than individual models furthermore time required for tuning the model was also less.

The article [16] describes the dropout prediction problem in EDX MOOCs using users behavior log data. The article relates to complete extraction of EDX data with classification model training including feature engineering and data preprocessing. The different model showed accuracy as logistic regression having accuracy 65%, support vector machine with 65% accuracy, random forest and gradient boosting decision tree with accuracy 85% and 88% respectively. Therefore, it concluded that ensemble models show a better prediction result than individual models. Feature selection can be performed in text analysis which can be reviewed by the article [17] . The article refers to feature engineering used for tweet classification which describes a large discrimination in various texts in a corpus. The article performed ten feature-based engineering techniques against the non-featured models to compare the results and compare the performance. On the other side, article [18] highlighted the use of machine learning classification models in field of largest cancer which is breast cancer. SVM was the classification model applied combining the different feature engineering attributes to derive accurate result. The performance of the applied model was evaluated using accuracy, sensitivity, specificity, positive and negative predictive values, which showed highest accuracy as 99.3% and promising results for breast cancer.

2.3 Study for Evaluation of models

The process of Evaluation plays a major role in implementing the model which can be reviewed by applying evaluation metrices for each model to check weather model performance towards the given dataset is effective or not. The article [19] reviews the performance of the model build for predicting the software defect. The evaluation metrices used are Area under curve (AUC) and precision recall curve. The best predicted model evaluated in AUC-ROC curve for different dataset shows that logistic regression plays important role in prediction. For highly skewed classification dataset precision recall plays the best role. The performance of different algorithms can be compared using accuracy, Area under curve (AUC) and Precision Recall (PC) can be studied from article [20] on the basis of average values and standard deviation. Cultural modelling is been used to perform to compare the models for organizational behavior pattern. The challenges faced in feature engineering is class imbalance occurrence which reduced the performance of certain models.

3. Methodology

For every research work there must be some methodology used to implement some calculation on data and play with it. As we can observe that large volume of data is increasing now-a-days to handle and derive informative result we must have certain methodology. The research project follows Knowledge Discovery Databases (KDD) methodology for processing the data and gather useful knowledge from the dataset. The dataset is analyzed and studied from different angles to discover knowledgeable relationships. This could make the process showed in figure 1 of decision making a bit easier for certain organization using such approach.

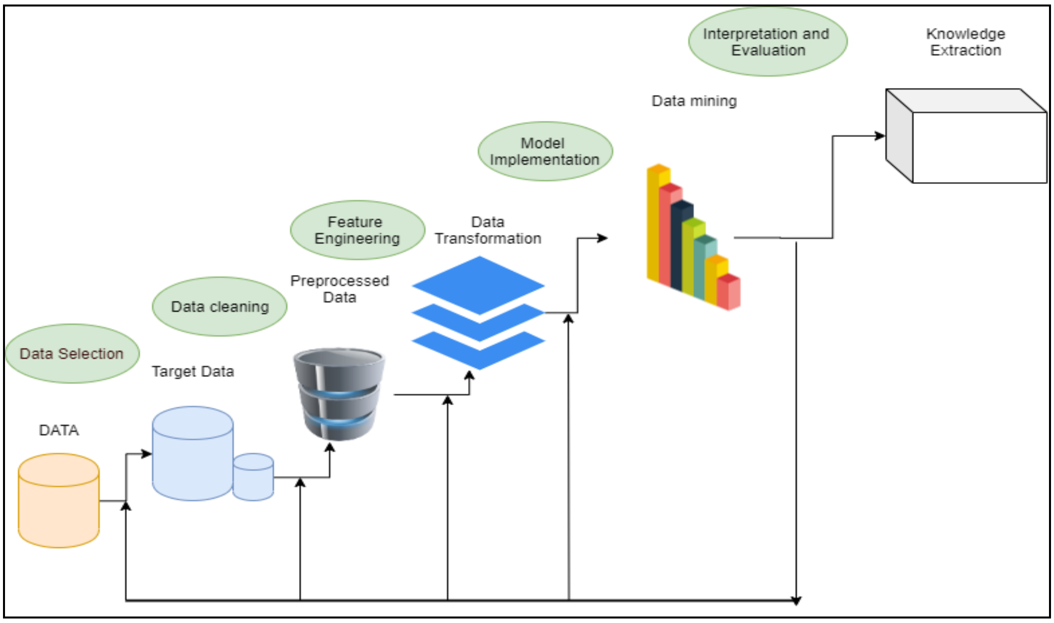


Figure 1: KDD Methodology

3.1 Dataset description

The section consists of dataset for the Austin Animal Center Shelter outcomes which is a comprehensive dataset consisting over 26 thousand inputs from October 2013 till March 2016 on an hourly frequency format. The study of research was considered as Austin is because it contains the world's largest no kill animal shelter with massive numbers of domestic animals whereas the data used is updated daily which helps in future. [4]The Dataset includes name, date of birth, outcome, animal type, sex, age at time of outcome, breed and color where outcomes ranges widely like adoptions, transfers, return to owner, death and euthanasia. The dataset was posted on Kaggle which was directly imported from Austin animal shelter website. Some of the attributes were directly used whereas some unwanted attributes were removed for further processing of the data variables. The attributes were classified based on outcome type to build the classification models.

[4] https://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-outcomes-and#aac_shelter_outcomes.csv

3.2 Data preprocessing

In advance to implementing any data mining models or machine learning algorithms it is necessary to clean the data and to remove irrelevant features and attributes to further process the data for fitting the model. The data for Austin animal shelter outcome consisted of lot of missing values and outliers in various attributes which were figured out using the null function (isnull). In date and time variable, cells having PM and AM timing structure were cleaned using the replace function in EXCEL. Using the library (tm) which is a text mining package was used to remove punctuation marks or more special characters in the whole data. These processes help in reducing noise from the data as it lacks on holding any useful information. The unwanted column like animal Id and date of birth were removed which did not convey any useful information. The functions were created as a control flow to pass. The functions were then used as an object to accomplish the required actions.

3.3 Feature engineering

Prior to building models and setting up the data to evaluate the performance of the model we must perform some feature engineering tasks on our preprocessed data. As we know our predictor is a categorical variable, we must group our other variables based on the predictor. The dataset was separated in granular format by creating function. The functions were created to convert age into number of days, convert time to period of day as morning afternoon or night-time, conversion of hybrid colors to common colors, creating dummy variables for colors. Constructing the following features includes-

1. **Cleaning name** – The name variable was having names and blanks, so the animals having no name were changed to unknown and a new variable as name status is included in the dataset having 1 and 0 for animals having name and unknown.

2. **Cleaning Sex upon outcome** – the missing value in this variable was imported with most common outcome the column. Separated sex upon outcome and intact status as do different variable for dogs and cats. The package *caret* was used for separation of the columns into two different columns.

3. **Cleaning datetime** – The date and time variable consisted of date and time in a single column which were separated as date and time of period separately. New predictor variable as year, month, weekday and time period were created and the date was duplicated from the datetime column using the mutate function installed by *dplyr* package. The function convert time to period was used to derive time when was the animal brought to animal center.

4. **Cleaning Breed**– The breed variable consists of original breed and mix breed which was separated by creating the new predictor variable as isMIX as 1 for mix and 0 for original breed.

5. **Cleaning Age upon outcome** – The variable age upon outcome was having age of animals in weeks and years which were converted to days by creating a new variable age in days using the *separate* function. The function convert age into number of days was used here. The age having 0 days were changed to NA values.

6. **Imputing NA values in Ages** – For imputing NA values PMM (predictive mean matching) method was used to validate imputed values are acceptable using set seed method to develop random variables. The mice (multivariate imputation by chained equation) algorithm was used which is capable of imputing mixes of binary and categorical variable. This process was done by creating a new data frame as mice.

7. **Categorizing colors** – [5]The color variable has single color or multi-color separated by '/' which was separated as color1 and color2 and further using function convert to common color and creating dummy variables six basic color predictors were formed as 'is black' , 'is gray', 'is brown' , 'is white' and ' is multi'.

8. **Remove unwanted variables** – Removed outcome subtype, name, breed date, time color which does not convey any information after feature engineering process and separation.

3.4 Data mining-

Data mining is the process where large amount of dataset can be transformed to derive informative patterns. In this research project we classified the predictor variable outcome type using the different supervised learning algorithms to correctly predict and classify which animals are likely to be classified as per the predicted outcomes as Adoption, died, euthanasia, transfer and return to owner. Cross validation is a statistical method is used to measure the skill of classification models by reducing errors which is a resampling procedure. K- fold cross validation is a process in which parameter k refers to number folds to be formed for the training data passed for cross validation with certain iteration. It is an approach that results in less biased and less optimistic estimation of model's skills with simple holdout approach of split. The project describes 4 different classification models as logistic regression (LR), Neural Network (NN), XGboost (XGB) and random forest (RF). The related work done by different researchers were the baseline for the models implemented.

3.5 Evaluation metrics-

In Process of KDD evaluation method is used to enhance mining algorithm. The preprocessed dataset is separated into two distinct set as training and testing in the ratio of 80:20. Hyperparameter tuning is being performed for the dataset to get the optimum results for algorithms used as neural network, XGboost, logistic regression and random forest. For

[5] https://cran.r-project.org/web/packages/colorSpec/vignettes/colorSpec-guide.html

evaluating performance, we used different metrices as: Accuracy, logarithmic loss, sensitivity and specificity.

4. Design specification-

The research project design is divided in 3 Tiers design Architecture namely-

1. Data persistent layer
2. Application layer
3. Client layer

Data persistent layer denotes the first stage of the project as the collection of data and Preprocessing of the data using feature engineering to create the monitoring data. Application layer consist of major part of the project where different classification models are being implemented using the R programming language. The third layer is client layer which represents the evaluation and performance of models on monitoring data to derive information for the Animal center in Austin to collect information and improve current work for animals.

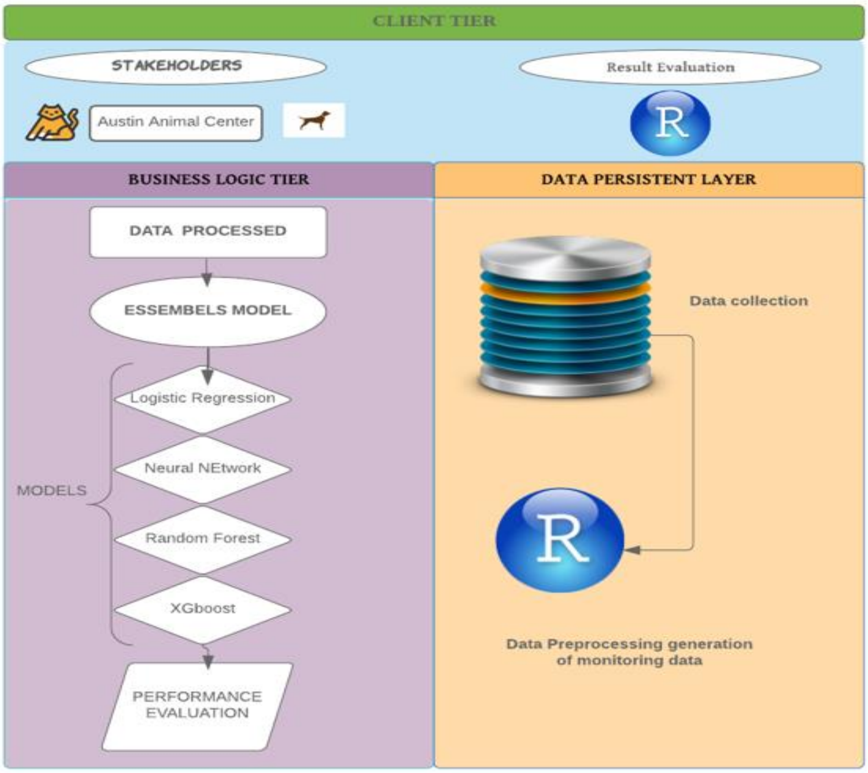


Figure 2: Three Tier Design Architecture

5. Process flow-

This section illustrates the basic flow of how the project is being implemented using different color which can be estimated from figure 3. The data illustrates different categorical variables which is being stated in the process flow. As the process of data preprocessing which is further extended to data cleaning, feature engineering and vectorization which is different approach encapsulated in the project marked with blue color boxes. The process of differentiating each categorical variable and adding predictive carriable is different and simple approach which can is used in the Data preprocessing. Further the process states the splitting of data as taking and testing and building classification models by the means of k-fold cross validation to avoid overfitting and to run the models smoothly. The whole process is being evaluated using the test data using evaluation metrics.

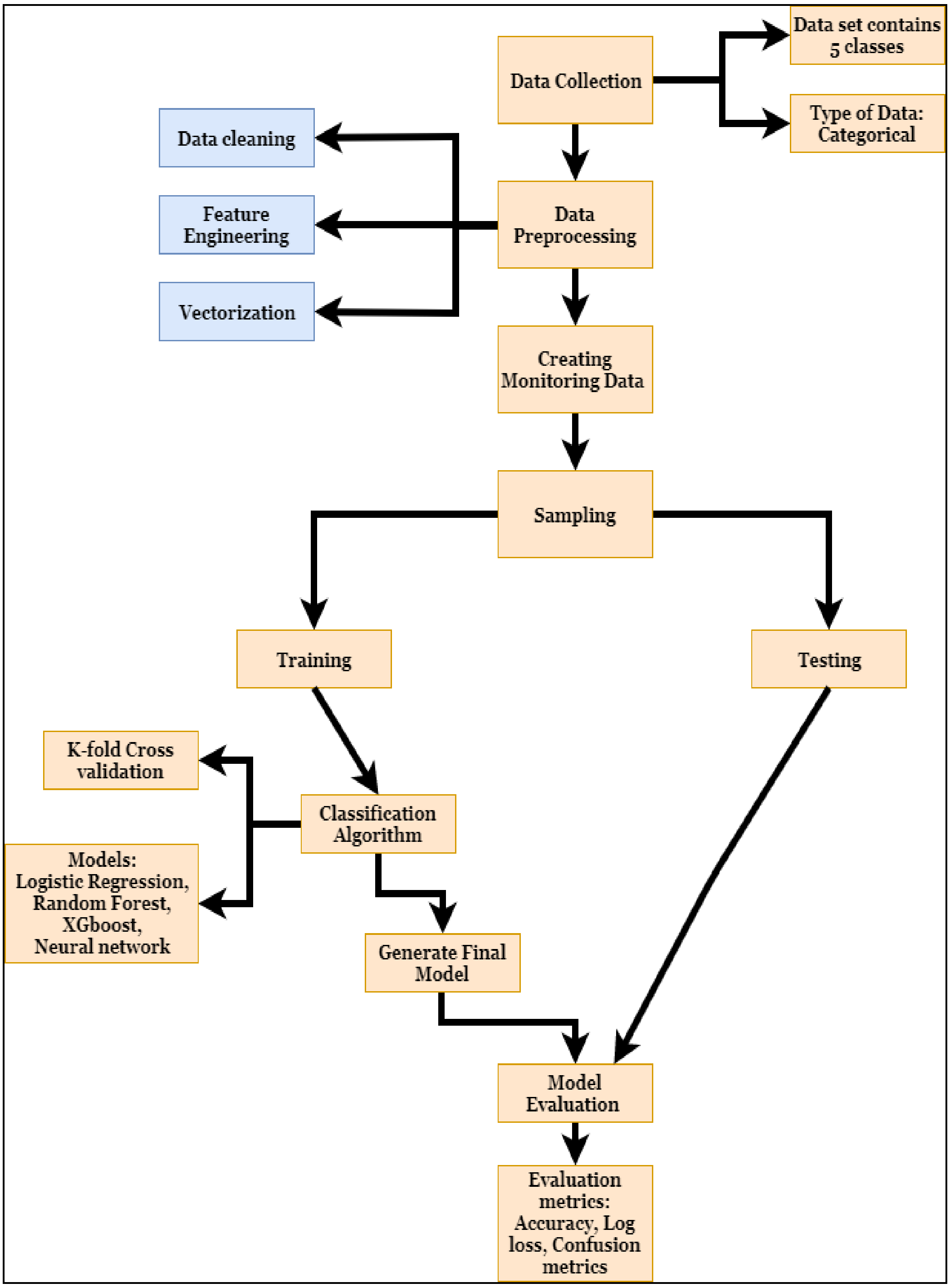


Figure 3: Process Flow of Research Project.

6. Implementation

This section summarizes the detail study to create and effective classification model, which were built to determine for Austin shelter to determine which animals are likely to be euthanized as well as certain trends in what feature increases the chances for Adoption which is used on RStudio environment which deals with Integrated Development Environment (IDE) using R codes. It is an open source and free platform. For implementing models' different packages and libraries were installed. Packages like *Tidyverse, Caret* and *Mice* were used for cleaning the dataset properly

The libraries were executed to perform the implementation process smoothly. Functions for cleaning the data were created to develop a different approach in the process of cleaning the dataset to fit the model. The data set was imported in the form of CSV using the *read.csv* function for the research. Each column was analyzed in order to perform the dataset to create monitoring data for prediction outcome for Austin animal shelter data. Each column was cleaned as per requirement and special characters were removed. The different approach like name was changed to name status by adding a new predictive column and animals having no name were changed to unknown. Columns like color breed were also cleaned by creating different predictive variable and later was removed. Variable date and time were differentiated as year, month, weekday and time period. The monitoring CSV was separated into training and testing set for model evaluation and implementation.

The 5-fold cross validation for train function of caret is used to evaluate using resampling and for the purpose of model estimation from training set of data. The model of cross validation is used for the models such as neural network, XGboost and random forest which can be estimated from *caret* function. In K-fold cross validation the training data passed is divided into small folds approximately of same size and iterated number of times. In each iteration certain data is taken as validation and fit for the rest of remaining data. This process is repeated for each iteration, so all the training data is set as validation in some of the iteration. The *traincontrol* function is used to as method to be cross validation, number of folds and summary as multiclass as there are different and multiple classes for prediction. The method explained is simple bootstrap samples which creates groups dynamically for the specified sample size which is 5 for the research project using the function *createmultifolds*.

6.1 Implementation of logistic regression model

Logistic regression is a process to fitting the curve $y = f(x)$ where y is categorical variable[6]. The basic use of the model is predicting y for given set of x predictors. This method is achieved by taking log odds considering each of the variable. The training data was created in design matrix using the *model.matrix* function. This matrix data frame was then split into training and testing data frames to perform the further process. The *Glmnet* library was installed to perform modelling. *cv.glmnet* is the main function to perform logistic regression

[6] https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/

for object *cv.fit*. The x for matrix in *Glmnet* and response matrix y is been created. In this model implementation we have used the family argument as multinomial which supports the prediction of classes up to 5 and an experiment argument as *type.multinomial* is stated as grouped. For fitting each fold of data, the argument parallel is kept as true with the specification of number of folds. The fitting curve is plotted as –

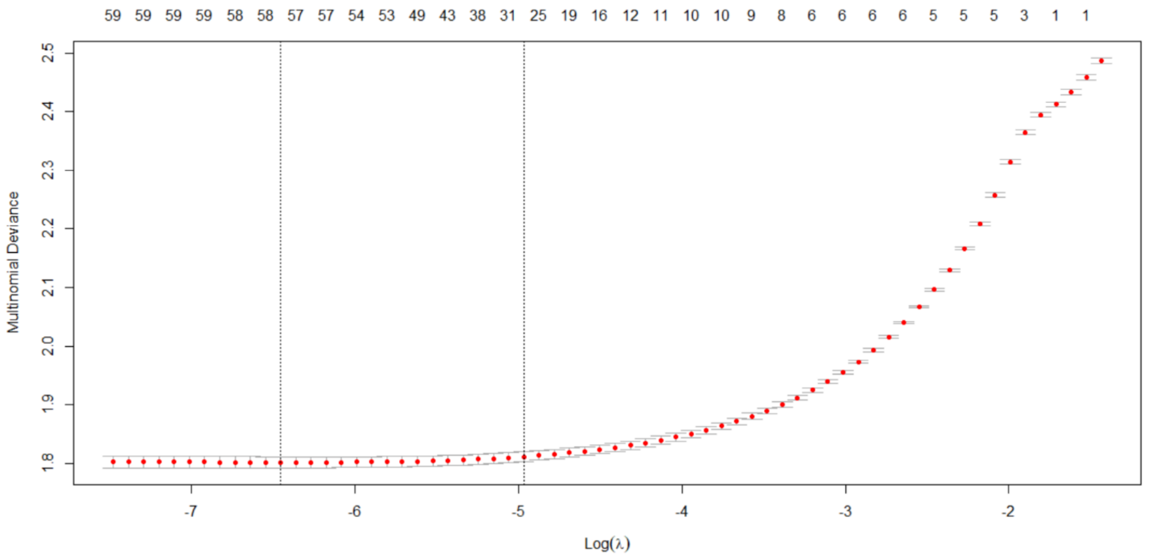


Figure 4: Cross Validation Curve for Logistic Regression

The above curve shown in figure 4 defines the cross-validation curve where the object fitted as cv.fit is being plotted. The curve shows left to right variable nonzero coefficients that is the data frames. The y axis denotes the percent of multinomial deviation in percentage for the value of lamda which is for default set to 100 but it stops at certain point as the deviation does not show much change in the next value of lamda. The line bordering each of the red dots are the error bars. The two vertical lines shows the two selected lamdas.

6.2 Implementing Neural Network

Neural networks are models which works basically on human brain's working principle which makes the computer capable for thinking. In this research project we use cross validation to build neural network model. This process can be used to test the quality of neural network. K-fold cross validation helps in choosing the best parameter as per assigned arguments for neural network. For training the model we first seed value to a random number which is used to generate random objects which can be reproduced. The object *nn.model* is used to define the neural network model. We use the train function to build the model with certain arguments. It sets up a tuning parameter for lots of regression and classification problems and performs resampling. Initialising with the dependent variable and the data to be used for executing the train function. The method to be used is *nnet* which specifies that classification model to be used. The arguments *tunelength* and *trcontrol* is used in arguments to define for the integer denoting number of values and *traincontrol* to be used as mentioned above. The metrics to be evaluated is defined as *logloss* which quantifies the accuracy of algorithm by penalising false classification. The number of iterations on whole training dataset is set to be 100 for different weights.

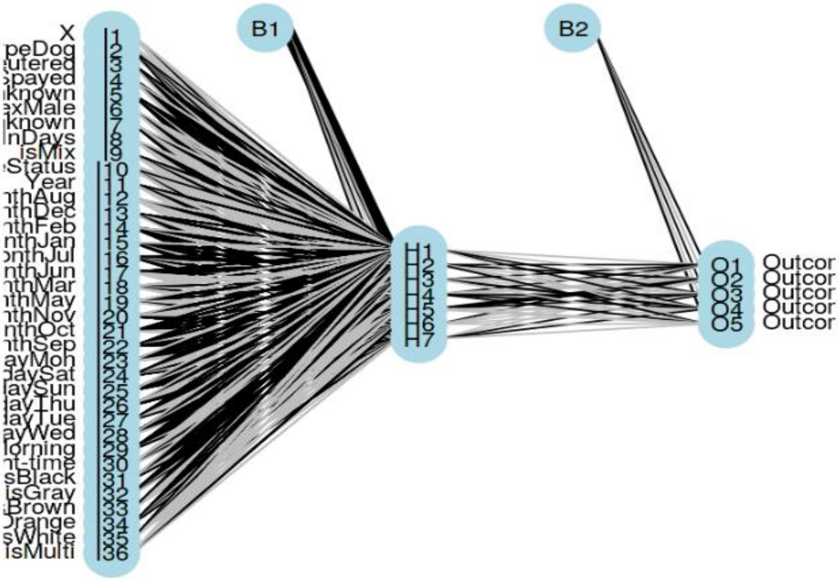


Figure 5: Implemented Neural Network (I- input layers, H- Hidden layers, O- output layers, B- Back propagation layers)

The above figure 5 shows the implementation of neural network for the monitoring dataset provided with 36 input layers, 7 hidden layers and 2 back propagation layers.

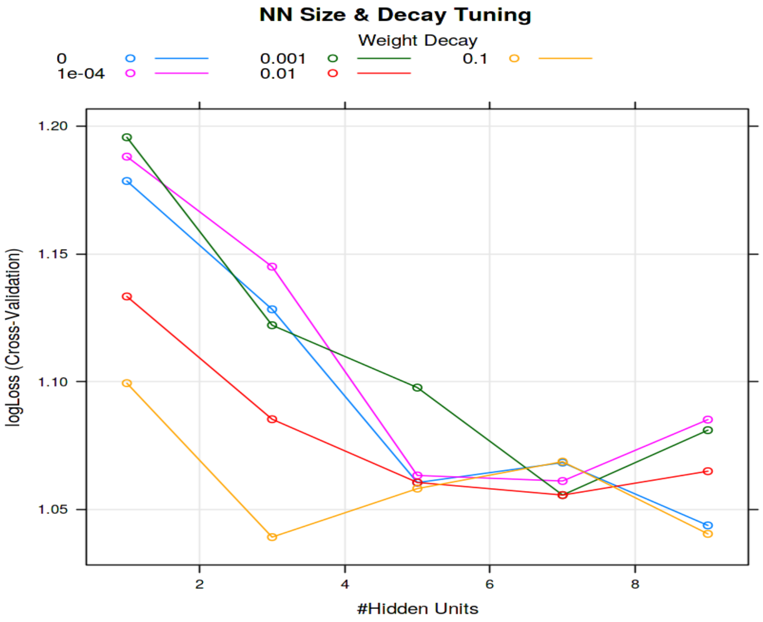


Figure 6: Learning Rate for Different Weights

As per the above figure 6 the learning rate for different decay weight can be evaluated. Using the log loss metrics, we can see that the is minimal for each for the weights due to reduce in overfitting of data. The learning rate on and average is assumed to be good learning rate.

6.3 Implementation of Random Forest

Random forest are ensemble learning algorithms basically used for classification purpose which operates by adding multitude decision trees. Decision trees are building blocks of random forest. For this research we used random forest as a classification algorithm to develop a different approach than applied models. For random forest we use function *train* to with the method argument as rf and new additional argument as ntree which is set to 500 and metric as logloss.

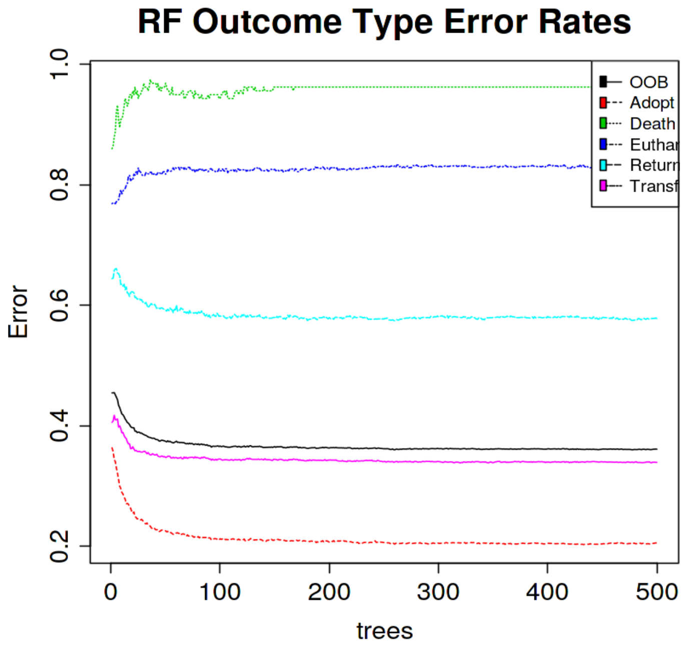


Figure 7: Out of Box plot For Random Forest

The above figure 7 shows that out of bag error and the error computed for different predictive class. We can observe that adoption and transfer predict very less error compared to death and euthanise. The OOB error is also very low so we can say that a lot of rows are been considered to build the random forest model.

6.4 Implementation of XGboost model

XGboost is a highly optimised distributed gradient boosting library designed for well organised approach towards machine learning algorithm to train and evaluate problems in accurate way[7]. For implementation of XGboost we import xgboost library. We will create Grid search which will tune the hyper parameters. Function *expand.grid* is used to define the grid search object. The booster parameter is used to create grid. The eta argument makes the

[7] https://xgboost.readthedocs.io/en/latest/

model more effective by minimizing the weights on each step where as max_depth argument used to control overfitting. Subsample argument is kept as 1 for default which denotes the fraction of sample that need to be randomly sampled. This grid created is further used to build the model. Setting the seed value to certain random number the train function is applied with method argument as *xgb_model* tuned with the created grid.

7. Evaluation

The research basic objective to build main models to accurately classify the outcome of Austin animal shelter center. This can be achieved by evaluating the boosting algorithms with the test data to know how well the they classify the outcome. The models are trained accurately to evaluate using Accuracy, sensitivity, specificity, logarithm loss and confusion matrix for each of the algorithm.

7.1 Accuracy

Accuracy is one of the metrics for evaluating classification models. The fraction obtained tells us how accurately our model is predicted. Accuracy can be calculated as

$$Accuracy = \frac{Number\ of\ all\ correct\ prediction}{Total\ number\ of\ prediction} \quad \text{.............} \ \mathbf{(1)}$$

Equation 1: Accuracy

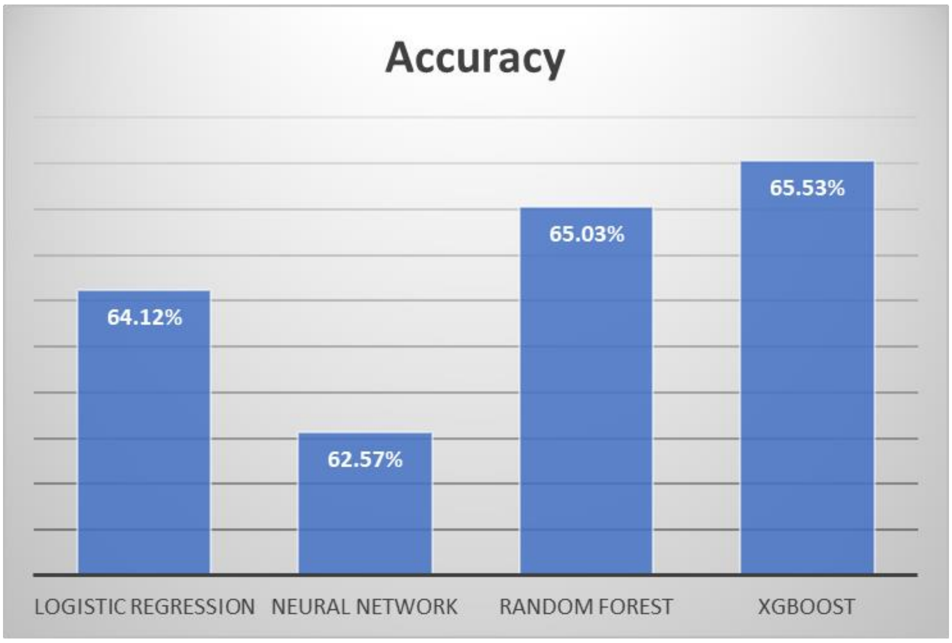


Figure 8: Comparison of Accuracies

The above figure 8 shows the evaluated accuracy for the implemented models. We can observe that XGboost has outperformed better than all the models with neural network having the least accuracy.

7.2 Logarithmic Loss

Logarithm loss is a measure for measuring the performance of the classification models in the probability between 0 and 1. The basic goal for the model to have high performance the logarithm loss should have value close to 0. Logarithm loss can be calculated using the below equation 2-

$$Logarithmic\ loss = \frac{-1}{N}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij} * \log(p_{ij}) \qquad \text{................ } \mathbf{(2)}$$

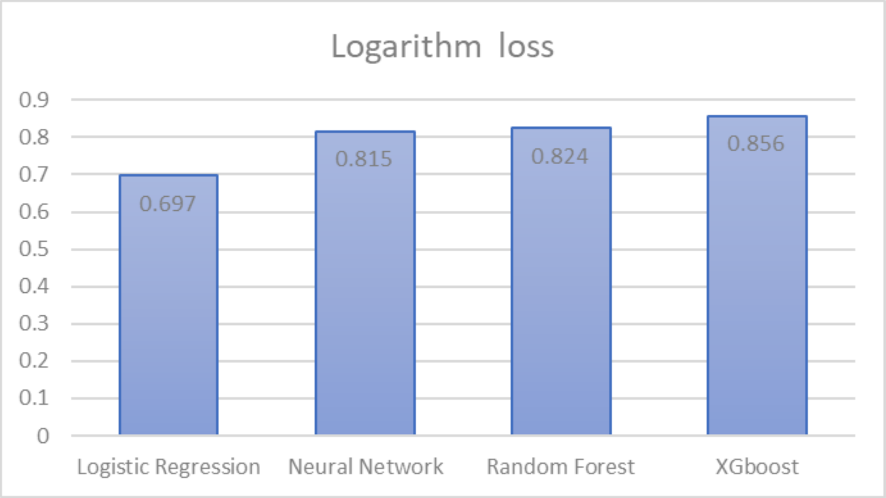


Figure 9: Logarithmic loss for different models

The performance metrics for all models were accurately similar other than logistic regression which showed best for classification with a score of 0.69 whereas XGboost showed a inaccurate results with score of 0.856 derived from the chart showed in figure 9.

7.3 Sensitivity

Term Sensitivity is defined as the proportion of actual cases which were positive were accurately predicted as positive. Sensitivity is also called recall. Sensitivity can be calculated as given below as equation 3-

$$Sensitivity = \frac{Number\ of\ correct\ positive\ prediction}{Total\ number\ of\ prediction} \qquad \text{............}\mathbf{(3)}$$

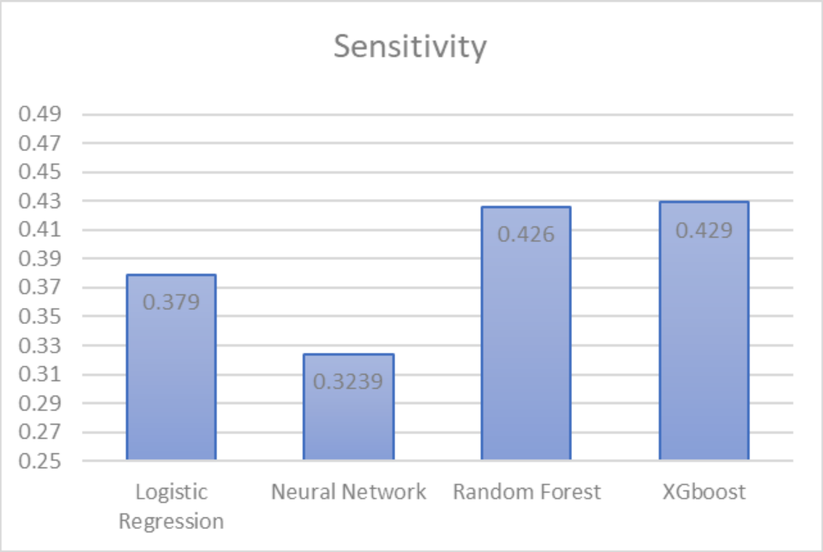


Figure 10: Comparison of Sensitivity

From the Above figure 10 as we can estimate that sensitivity of XGboost was higher than other models following random forest, logistic regression and neural network. The highest actual outcomes which were predicted truly as same outcome was performed well by XGboost.

7.4 Specificity

Term specificity is defined as the proportion of actual cases which were negative were accurately predicted as negative. Specificity can be calculated by using equation 4-

$$Specificity = \frac{Number\ of\ correct\ negative\ prediction}{Total\ number\ of\ prediction} \quad \text{.............} \ \mathbf{(4)}$$

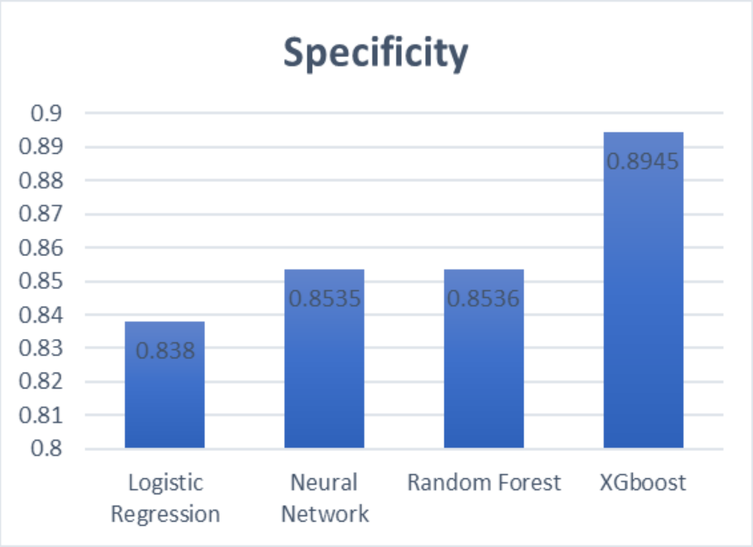


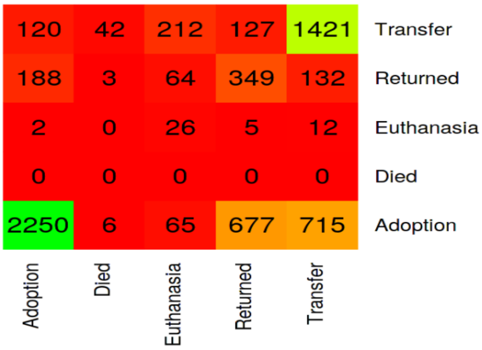
Figure 11: Comparison of Specificity

The above figure 11 shows comparison of different specificity values for different models. XGboost shows a pretty high value for wrong prediction in confusion matrix than the actual value. Logistic regression showed a lowest value compared to other models.

7.5 Confusion matrix

The performance of the classification algorithm can be tested based on confusion metrics. Accuracy cannot alone determine the performance of the model is accurate or not, whereas confusion matrix can determine the actual and predicted values for each of the classification model. Confusion metrics give us the idea how well the model is performing and where the errors are made in prediction. The vertical columns in the below confusion matrixes are the actual outcome and horizontal class are the predicted outcomes.

1. **Confusion metrics for Logistic regression-**

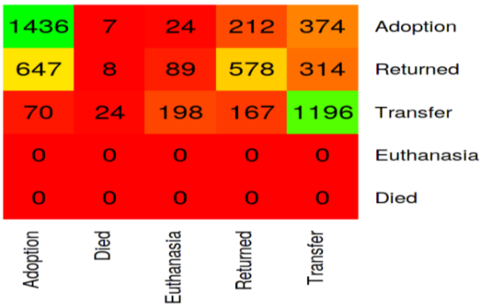


Adoption	Died	Euthanasia	Returned	Transfer	
120	42	212	127	1421	Transfer
188	3	64	349	132	Returned
2	0	26	5	12	Euthanasia
0	0	0	0	0	Died
2250	6	65	677	715	Adoption

Figure 12: Confusion matrix for Logistic Regression

The Above figure 12 shows confusion matrix for Logistic regression algorithm. The vertical class defined as actual values and the horizontal class is defined as predicted values. The confusion correctly classified 2250 as adoption among the total results for adoption and for rest of the outcomes adoption was predicted as transfer (715), returned (677), euthanasia (65) and died (6). The diagonal numbers from bottom left to extreme right are truly predicted whereas the other predicted values were some other class and predicted as different class.

2. **Confusion metrics for Neural Networks-**

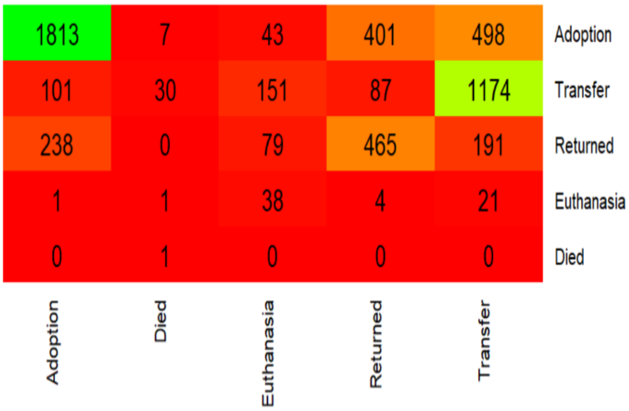


Adoption	Died	Euthanasia	Returned	Transfer	
1436	7	24	212	374	Adoption
647	8	89	578	314	Returned
70	24	198	167	1196	Transfer
0	0	0	0	0	Euthanasia
0	0	0	0	0	Died

Figure 13: Confusion matrix for Neural Networks

The above heat map visualization as figure 13 is the build confusion metrics for neural network model. The adoption rate for actual and predicted values was truly predicted as adoption as 1436 and for predicted died as 7, euthanasia as 24, returned as 212 and transfer as 374. For transfer class 1196 was correctly predicted as transfer and other values for prediction was wrongly predicted as different class which tells the performance of model from logarithmic values. The other class does not much of the difference in actual and predicted values.

3. Confusion metrics for XGboost-

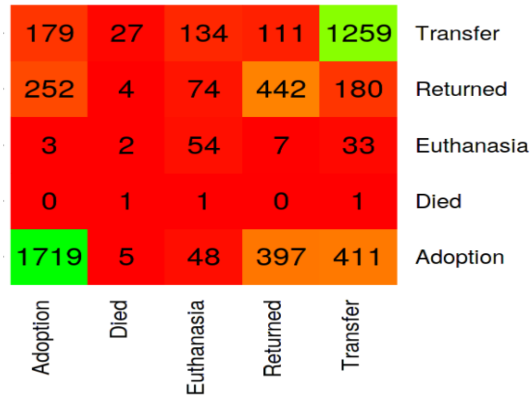


1813	7	43	401	498	Adoption
101	30	151	87	1174	Transfer
238	0	79	465	191	Returned
1	1	38	4	21	Euthanasia
0	1	0	0	0	Died
Adoption	Died	Euthanasia	Returned	Transfer	

Figure 14: Confusion matrix for XGboost

The above figure 14 shows confusion matrix for the XGboost model implemented in a heat map visualization. XGboost predicted the best values compared to other models. Most of the actual and predicted values for different outcomes were high and correctly predicted. For the class adoption 1813 were predicted correctly as adoption whereas the wrongly predicted class were transfer (7), euthanasia (43), returned (401) and transfer (498). Class transfer also showed much accurate result as 1174 as predicted transfer from 1543 outcomes, so the wrong prediction rate was much lower compared to other models.

4. Confusion metrics for Random Forest-



179	27	134	111	1259	Transfer
252	4	74	442	180	Returned
3	2	54	7	33	Euthanasia
0	1	1	0	1	Died
1719	5	48	397	411	Adoption
Adoption	Died	Euthanasia	Returned	Transfer	

Figure 15: Confusion matrix for Random Forest

The above figure 15 is confusion matrix build for Random forest. The truly predicted values for each of the class can be estimated as form the bottom left to the extreme right. For the outcome adoption it predicted 1719 correctly as adoption form 2580 values which estimated a large proportion of wrong prediction as per the actual class. For other classes the imbalance is a bit large due to low performance calculated with low accuracy.

8. Discussion

From the above results explained we are more likely to assume that out of the four algorithms implemented XGboost model showed pretty much accurate performance. The prediction analysis from the confusion matrix also predicted much accurately, for the actual values for the XGboost model. The classification model which gave the least accurate result was Neural Network with least accuracy and inaccurate results predicted from the confusion matrix. The results for Random forest and Logistic regression were better than Neural Network. K- fold classification derived resampling techniques to provide the best accurate results. From the above results obtained we can observe that classification was much accurate for logistic regression from logarithmic values, but the accuracy and confusion matrix gave best results for XGboost model. Therefore, we can say that ensembled models are more accurate than model on their own, but the results derived were time consuming.

The K-fold cross validation for the research project was 5 which can be increased to derive more accurate results in future, but the drawback of increasing the fold size will result in increase in variance in the sample data. K- fold cross validation improvisation classification by 3% then the generalized approach. The important feature that increase the rate of adoption can be predicted form the statics model. The major contribution of the research work is derived from the feature engineering performed to fit the models using the vectorization approach. the predictive results can be further used for the center to further update the program and amenities to be improved as per the number of animals to be adopted or transferred over the years.

9. Conclusion and Future work

The Research question basically concludes the idea of increasing the idea of maximizing the rate of adoption by building models and minimizing the rate for the euthanasia, which is been executed in this research project. In this research, we presented an implementation of four classification models, to predict and classify different outcome, to increase the rate of adoption in future or Austin Animal center. Moreover, the creation of monitoring data to fit the model using vectorization approach which provided pretty much satisfactory results. K-fold cross validation was used with hold out approach to resample the dataset. The classification process considered all the valuable features that could improve the rate of adoption. As the dataset used for implementing the project is updated daily so the best model, as per convenience can be used to derive the predictive results in future. For future work the models can be trained differently using fuzzy logics to adapt more accurate results. The project therefore concludes to be a never-ending work for the animal centers providing shelters to animals. This work can help in boosting the sectors of veterinary shops and increase the demand for veterinary doctors due to increase rate of adoption. The research work can be used by pet adopting software to make easy for the owner to adopt the animal with no second thought by inculcating the different features as per wish. This project can be further used by different animal care agencies to make the rate of adoption increase or decrease the death rate of animals.

References

[1] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Jan. 2015.

[2] X. Zhou, K.-Y. Liu, and S. T. C. Wong, "Cancer classification and prediction using logistic regression with Bayesian gene selection," *J. Biomed. Inform.*, vol. 37, no. 4, pp. 249–259, Aug. 2004.

[3] I. Kurt, M. Ture, and A. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 366–374, Jan. 2008.

[4] S. Jain and H. Kaur, "Machine learning approaches to predict basketball game outcome," in *2017 3rd International Conference on Advances in Computing,Communication & Automation (ICACCA) (Fall)*, 2017, pp. 1–7.

[5] M. T., D. Mukherji, N. Padalia, and A. Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)," *Int. J. Comput. Appl.*, vol. 68, no. 16, pp. 11–15, 2013.

[6] Y. Xing, J. Wang, Z. Zhao, and andYonghong Gao, "Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease," in *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, 2007, pp. 868–872.

[7] J. Chen, G. Xi, Y. Xing, J. Chen, and J. Wang, "Predicting Syndrome by NEI Specifications: A Comparison of Five Data Mining Algorithms in Coronary Heart Disease," in *Life System Modeling and Simulation*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 129–135.

[8] Y. Pang, N. Judd, J. O'Brien, and M. Ben-Avie, "Predicting students' graduation outcomes through support vector machines," in *2017 IEEE Frontiers in Education Conference (FIE)*, 2017, pp. 1–8.

[9] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Stat. Comput.*, vol. 21, no. 2, pp. 137–146, Apr. 2011.

[10] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-

one-out cross validation," *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, Sep. 2015.

[11] T. Kawamura, T. Toma, and K. Takano, "Outcome prediction of software projects for information technology vendors," in *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2017, pp. 1733–1737.

[12] A. Khemphila and V. Boonjing, "Heart Disease Classification Using Neural Network and Feature Selection," in *2011 21st International Conference on Systems Engineering*, 2011, pp. 406–409.

[13] A. Srivastava, A. Mahmood, and R. Srivastava, "A Comparative Analysis of SVM Random Forest Methods for Protein Function Prediction," in *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, 2017, pp. 1008–1010.

[14] J. Drgoňa, D. Picard, M. Kvasnica, and L. Helsen, "Approximate model predictive building control via machine learning," *Appl. Energy*, vol. 218, pp. 199–216, May 2018.

[15] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," in *SoutheastCon 2016*, 2016, pp. 1–6.

[16] J. Liang, C. Li, and L. Zheng, "Machine learning application in MOOCs: Dropout prediction," in *2016 11th International Conference on Computer Science & Education (ICCSE)*, 2016, pp. 52–57.

[17] J. D. Prusa, T. M. Khoshgoftaar, and D. J. Dittman, "Impact of Feature Selection Techniques for Tweet Sentiment Classification," *Twenty-Eighth Int. Flairs Conf.*, Apr. 2015.

[18] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3240–3247, Mar. 2009.

[19] S. A. Khan and Z. Ali Rana, "Evaluating Performance of Software Defect Prediction Models Using Area Under Precision-Recall Curve (AUC-PR)," in *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)*, 2019, pp. 1–6.

[20] Xiaochen Li, Wenji Mao, Daniel Zeng, Peng Su, and Fei-Yue Wang, "Performance evaluation of classification methods in cultural modeling," in *2009 IEEE International Conference on Intelligence and Security Informatics*, 2009, pp. 248–250.