Application of Data Mining Algorithms to Predict Adoption Rate of Shelter Pets

Nandita Viswanath
Purdue University
viswanan@purdue.edu

Mohit Gupta
Purdue University
gupta440@purdue.edu

ABSTRACT

Nearly 1.5 million pets are euthanized every year in shelters because of the lack of resources and inability to find new owners. This project aims to utilize data mining algorithms on a data-set consisting of the pet adoption listings to predict the adoption rate of shelter pets. This information can be used by shelters to forecast the resource requirement for pet care as well as to identify key influencers in the adoption advertisement that impact the adoption rate. This problem is a multi-classification problem consisting of five distinct classes corresponding to the adoption speed of the pet. Three data mining algorithms are applied on the dataset, namely, Support Vector Machine, Decision Tree and Artificial Neural Network. The performance of the three models were compared and it was found that the neural network gave the best F1 score of 0.44, Decision Tree followed with an F1 score of 0.37 and SVM gave a score of 0.33. As an additional analysis, the multi-classification problem was extended to a binary classification problem aimed at predicting whether the pet would be adopted or not. In this case, the neural network once again gave the best F1 score of 0.78.

1. PROJECT OVERVIEW

The project team comprises of Mohit Gupta and Nandita Viswanath. The objective of the project is to predict the adoption rate of shelter pets based on a dataset of pet listings and compare the performance of Support Vector Machine, Decision Tree and Artificial Neural Network. The dataset is obtained from a Kaggle competition and provided by PetFinder.my [7], a shelter based in Malaysia.

2. INTRODUCTION

People purchase or adopt pets with the desire for a long term companion. However, after adopting a pet, many owners leave the pet in a shelter. The reasons for doing this vary - some people feel they do not have enough time for the pet, some think that the pet grew bigger than expected, some due to moving etc [4]. In the US, statistics show that a total of 6.5 million pets are left in shelters every year [6]. Although the shelters do their best to accommodate pets that are dropped off, it is often very difficult because of the resources required to care for each pet. Moreover, pets that have lived with families for years start developing aggressions and health issues due to fear and the new environment [3]. If shelters are not able to find a home for the pet or the pets fall too sick, the pets are euthanized. Statistics show that nearly 1.5 million pets that arrive at shelters are euthanized every year [6]. In some crowded shelters, the pet is euthanized if he/she cannot be adopted in 72 hours because the shelter workers are unable to predict or allocate enough resources to care for the pet [1].

Through this project, the aim is to use data mining algorithms on a dataset of pet adoption listings to predict the adoption rate for dogs and cats in a shelter. Using this information, shelters will be able to forecast the resources required for pet care and reduce the number of pets that they are forced to euthanize. Analyzing the description posted for each pet listing and the number of pictures or videos uploaded, we can try to identify key factors in the advertisement that have a positive effect on the adoption rate. Further, by understanding this information, awareness can be created to encourage people to adopt pets from shelters rather than purchase them from a pet store.

The report is structured as follows: Section 3 describes the exploration of the dataset and presents key visualization results, Section 4 delineates the design and implementation of the chosen data mining algorithms along with details about evaluation measure i.e. F1 score. Section 4 also contains a detailed description on preprocessing on the data set. Section 5 is the results section. This section has learning curves for all the three models and performance comparison of the algorithms. Finally, section 6 concludes with the findings and future scope of this work.

3. DATA EXPLORATION

Data exploration was performed in two steps - the first step focused on understanding the data and the feature types. Data visualization was performed as the second step to identify apparent trends in the dataset and apprehend the impact, if any, of the different features on the adoption rate.

3.1 Understanding the Dataset

The PetFinder data includes numeric, categorical, text and image data which can be used to predict the label value, speed of adoption. All categorical features are label encoded

 ISBN X-XXXXX-XX-X/XX/XX.
DOI: 000000

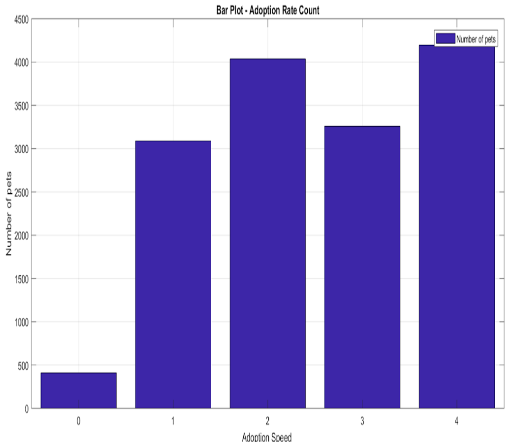

Figure 1: Bar Plot: Adoption Speed

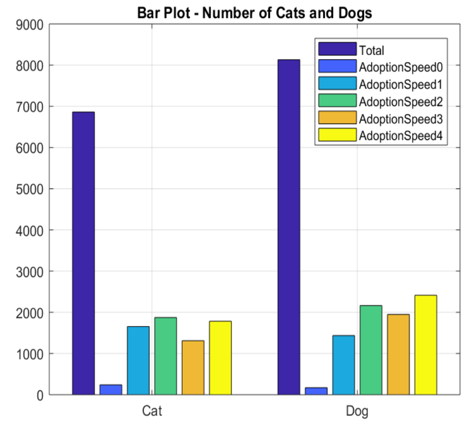

Figure 2: Bar Plot: Adoption Speed vs Pet Type

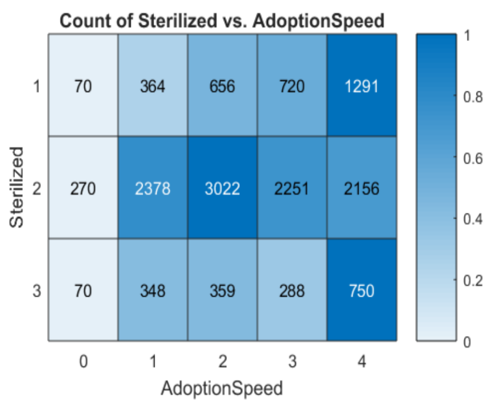

Figure 3: Adoption Speed vs Pet Sterilization

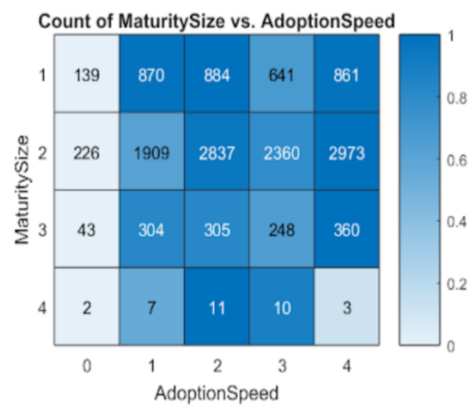

Figure 4: Adoption Speed vs Maturity Size

in the dataset and the description for each of the features is obtained from the source of the data [7]. The features include a Pet ID specific to every pet. The speed of adoption varies from 0 to 4: 0 - the pet is adopted the very day it was listed, 1 - pet adopted between 1 and 7 days after being listed, 2 - pet is adopted between 8 and 30 days, 3 - pet is adopted between 31 and 90 days and 4 - pet is not adopted after 100 days. The types of pets include two categories: dogs and cats. Other features include the age and breed of the pet. The breed of the pet is further divided into two categories: if the second breed is 0, then the pet is a single breed pet and if not, it is a hybrid. The size of the pet ranges from small to large. The state of vaccination, sterilization and deworming is listed as yes, no or unsure. The health of the pet could either be healthy, minor injury or serious injury. In addition, the dataset includes quantity, fee and state location of the pet in Malaysia. Finally, the dataset includes a text description of the pet. This text description is later preprocessed in this analysis using TF-IDF and Word2Vec embeddings.

3.2 Data Visualization

The features in this dataset are visualized using bar plots, bivariate histograms, cumulative distribution function plots and heat maps. The plots were created using MATLAB. The following features with discrete values are treated as categorical variables: adoption speed (label), type, gender, color, breed, maturity size, fur length, vaccinated, dewormed, sterilized and heath. Age, quantity and fee are considered as numeric features since they are not discrete.

Bar plots can be used to visualize the distribution of number of pets in different adoption speed categories as seen in Figure 1. From the bar plot, it is observed that the highest number of pets fall in category 4, ie. around 4200 pets out of the total 14000 pets are not adopted even after 100 days of being listed. This is closely followed by adoption speed 2 (around 3250 pets), 3 (around 3100 pets) and 1 (450 pets) respectively. Lesser than 500 pets fall in the adoption speed 0 category implying that few pets are adopted on the same day they are put up for adoption. The bar plot confirms that there is no one-sided trend describing number of pets in different adoption speed categories. This data can be sliced further to look at adoption speeds for different pet types. This will help understand if the pet type bears an apparent effect on speed of adoption.

The stacked bar plot in Figure 2 divides the previous plot by pet type. From this bar plot, it is seen that the number

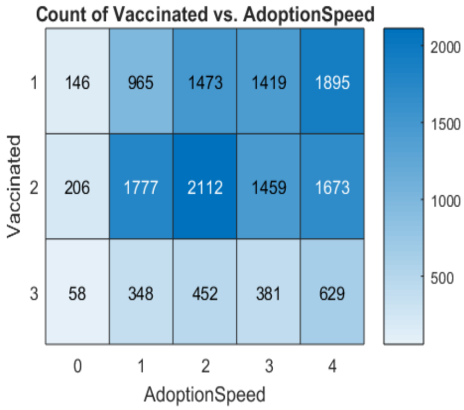

Figure 5: Adoption Speed vs Pet Vaccination

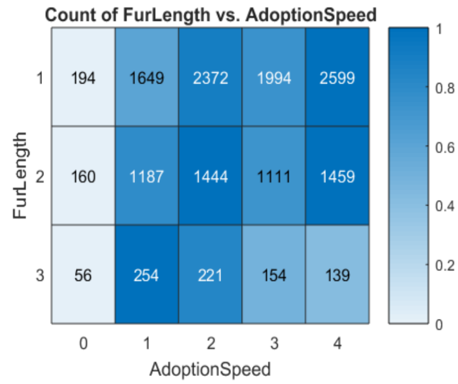

Figure 6: Adoption Speed vs Fur Length

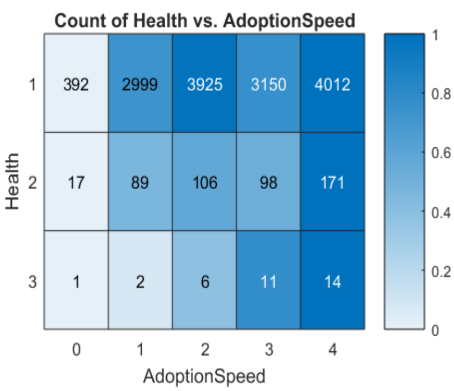

Figure 7: Adoption Speed vs Pet Health

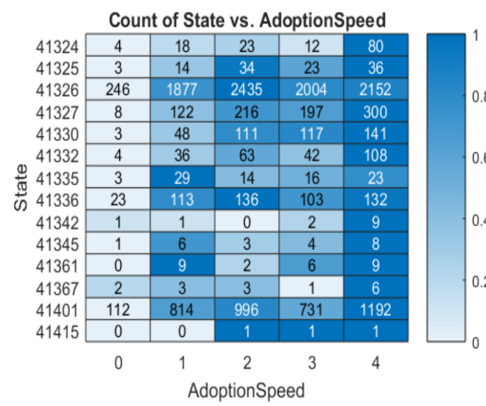

Figure 8: Adoption Speed vs State Location

of type 2 (dogs) pets are more than the number of type 1 (cat) pets. The bar plot for dogs shows a trend similar to the bar plot for the entire dataset with the highest number of dogs not being adopted even after 100 days of being listed. However, it is seen that the highest number of cats are adopted at adoption speed 2 (8 to 30 days) followed by 4 and 1 closely. The number of pets in the speed 0 category are the lowest for both cats and dogs but more cats than dogs are adopted on the same day they are left at the shelter, though there are more dogs at the shelter. This shows that a greater proportion of cats are adopted between 1 and 30 days of being listed than dogs.

A powerful method to visualize relationship between two categorical variables in using a 2-D heat map. The colorbar in the heatmap is normalized by the column. The color gradient is interpolated to represent the number of pets that fall within different categories. Column normalization will help to understand if the adoption speed trend varies for different row categories. In the heatmap shown in Figure 4, it is seen that the highest number of pets fall in the maturity size 2 (medium-sized) category followed by 1, 3 and 4. For pets that are small and medium sized (maturity size 1 and 2), almost equal number of pets are adopted at adoption speed 1 and 4 followed by 3 and 1. While there is no apparent trend that is deciphered from this plot, it is seen that 67 percent of the pets left at the shelter are medium-sized. Similarly, looking at sterilization of pets in Figure 3, it is seen that most pets that are adopted are not sterilized and are also adopted at higher adoption speeds. A greater proportion of pets that are sterilized are not adopted after 100 days than pets that are not sterilized. The color coding in the first two rows of the heatmap reflect this trend.

Looking at the 2-D heatmap for adoption speed and pet vaccination in Figure 5, it is observed that vaccination does not seem to be a major factor of consideration in pet adoption. 5898 pets left at the shelter are vaccinated (vaccinated 1) and 7227 pets are not vaccinated (vaccinated 2). It is observed that a greater proportion of pets that are not vaccinated are adopted at a greater speed than pets that are vaccinated. 32 percent of the pets that are vaccinated are not adopted after 100 days while 29 percent of pets that are not vaccinated are adopted at adoption speed 2. Although this does not necessarily imply that non-vaccinated pets are preferred (due to the lesser number of vaccinated pets in the dataset), it is likely that the vaccination of a pet does not directly favour the adoption speed.

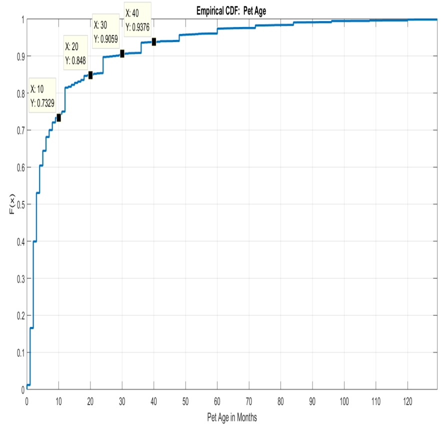

Figure 9: Age: Cumulative Distribution Function

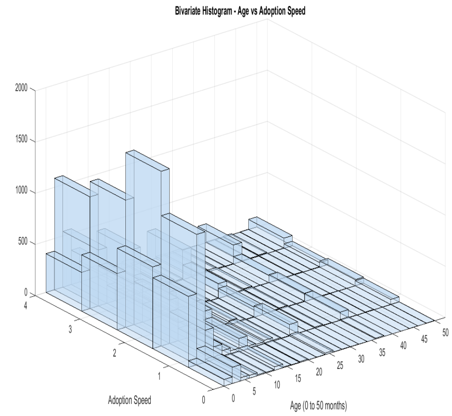

Figure 10: Speed vs Age: Bivariate Histogram

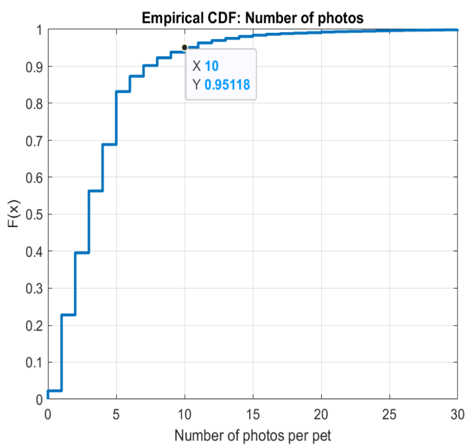

Figure 11: Photo Amount: Cumulative Distribution Function

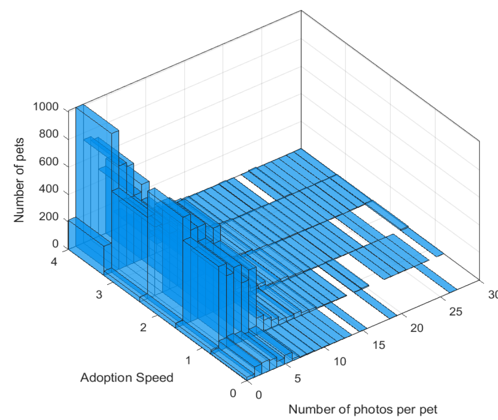

Figure 12: Speed vs Photo Amt: Bivariate Histogram

To investigate the impact of fur length on adoption speed, the heatmap in Figure 6 is used. It is observed that a higher proportion of pets with fur length 3 (longer fur) are adopted at a faster rate than any pets with shorter fur (1 and 2). For pets with short fur (fur length 1), the highest proportion of pets are not adopted after 100 days. From the health vs adoption speed heat map, it is observed that most pets are left at the shelter in good health and the adoption trend is similar to the overall population trend as seen in Figure 7. For pets that have a minor injury or a serious injury, the highest proportion of pets are not adopted even after 100 days. From the state vs adoption speed heatmap in Figure 8, it is observed that 60 percent of the pets are listed in zip code 41326 and 25 percent of the pets are listed in zip code 41401. In zip code 41326, the highest proportion of pets are adopted between 8 and 30 days of being listed the shelter while in 41401, the highest proportion of pets are not adopted even after 100 days of being listed.

Out of the 14993 pets, there are 188 different breed types in the dataset. Breed 1 has 176 different types of breeds and Breed 2 has 135 different types of breeds. Pets with a non-zero breed 2 breed type are hybrid pets. Out of the 176 types in breed 1, 13114 pets fall within the top 15 types of breed 1 breed types and 14296 pets fall within the top 15 types of breed 2 breed types. Further, it is seen that almost 6000 pets out of the 14993 are of breed type 307 (mixed breed). This is followed by breed type 266 (Domestic Short Hair) and 265 (Domestic Medium Hair) while other breed types constitute less than 3.3 percent of the total number of pets. In breed 2, it is seen that more than 66 percent of the pets are of breed type 0 (single breed).

The age of pets is treated as a numeric feature and to understand the distribution of age of pets, a cumulative distribution function (CDF) plot is generated. Using the CDF plot, inferences such as the proportion of pets below a certain age can be obtained. Using observations from the CDF plot, a bivariate histogram can be created to decipher the relationship between adoption speed and pet age.

From Figure 9, it is inferred that 73.29 percent of the total number of pets are lesser than 10 months old. 11.5 percent of the pets are between 10 and 20 months old and only 6.2 percent of the pets are greater than 40 months old. This seems to suggest that older pets are less likely to be left at the shelter. Since majority of the pets are lesser than 50 months old, only pets that are lesser than 50 months old are visualized in the bivariate histogram to understand the relationship between pet age and adoption speed. In Figure 10,the X-axis (to the right of the origin) bin size is 2 months. The Y-axis (to the left of the origin) consists of one bin for each level of adoption speed. It is seen that for pets between 0 and 4 months old, the most frequent adoption speed is 2 (between 8 and 30 days after being listed), followed by 1,4 and then 0. For pets that are between 4 and 6 months old, more pets are adopted at speed 3 (adopted between 31 and 90 days) followed by 4, 2, 1 and 0. As pets become older, the trend suggests that pets are adopted at slower speeds (3) or not adopted even after 100 days (4).

The number of photos available for each pet is another numeric feature that is used in the model. Out of 14993 pets, 413 pets had no photos and the rest had at least one photo along with the listing. Looking at the CDF distribution in Figure 11, it is understood that 95 percent of the pets have lesser than 10 photos in their listing and the maximum number of photos per pet is 30. In the bivariate histogram visualization in Figure 12, the bin-size for the x-axis is one: one bin corresponds to one photo per pet. It is seen from the bivariate histogram that the number of photos per available in the listing does not directly affect the adoption speed of the pet. As seen in the adoption speed bar plot, the trend continues to show that the highest proportion of pets are not adopted after 100 days of being listed.

4. METHODOLOGY AND ALGORITHMS IMPLEMENTED

4.1 Algorithms

In this work, adoption speed prediction is modelled as a classification problem in supervised learning. Supervised learning is one of the most popular machine learning methods. Supervised learning [9] uses labeled data to train models. The goal of this type of training is to classify the test data as one of the possible labels i.e. adoption speed 0 to 4 in the petfinder dataset on the basis of feature vectors. Unsupervised learning, on the other hand, uses unlabeled or untagged data to perform the learning task. In this work, our primary focus is on supervised learning and we use such methods to predict adoption speed. We implemented 3 data mining algorithms and used them to predict adoption speed in petfinder dataset. The algorithms are Support Vector Machine (SVM), Decision Trees and Artificial Neural Network (ANN).

An artificial neural network [2] is a network of simple elements called artificial neurons, which receive input, change their internal state (activation) according to that input, and produce output depending on the input and activation. Artificial Neural Networks have been around for many decades and have been gaining and losing the favor of research community. Application of Deep Neural Networks for the solution of numerous data mining problems is a relatively popular area of research. We tested with a few settings of the number of hidden layers and neurons in each layer, and then settled for the Artificial Neural Network that gave us high accuracy and F-1 scores - which for us consisted of two hidden layers, having fifteen hundred and five hundred neurons respectively.

In this work, we focused on comparing ANN based approach with traditional machine learning methods notably Support Vector Machine (SVM) model and Decision Tree Classifier. After preprocessing the data, we split the dataset with 70 percent of dataset serving as training dataset and the remaining 30 percent serving as test dataset. We opted to train all models on training dataset and then tested/evaluated the models on testing dataset.

4.2 Evaluation

For the purpose of evaluation, we use the metrics defined below by Joshi [5]:

1. **True Positives (TP)** - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

2. **True Negatives (TN)** - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

3. **False Positives (FP)** - When actual class is no and predicted class is yes.

4. **False Negatives (FN)** - When actual class is yes but predicted class in no.

5. **Accuracy** - Accuracy is a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model.

   Accuracy = TP+TN/TP+FP+FN+TN

6. **Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. Precision = TP/TP+FP

7. **Recall (Sensitivity)** - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

   Recall = TP/TP+FN

8. **F1 score** - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it is better to look at both Precision and Recall.

   F1 Score = 2*(Recall * Precision) / (Recall + Precision)

   To evaluate and compare the three models, we used micro F1 score as the metric. We used 5-Fold Cross Validation on the training set. The average F1 score values are given in the Results section along with the learning curves for all three algorithms.

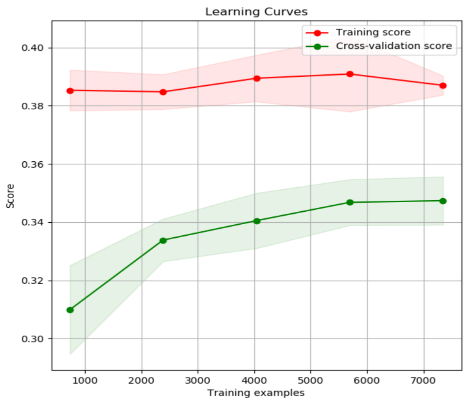

Figure 13: Learning Curve SVM

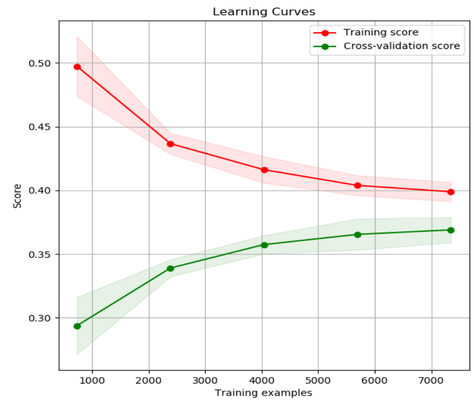

Figure 14: Learning Curve Decision Tree

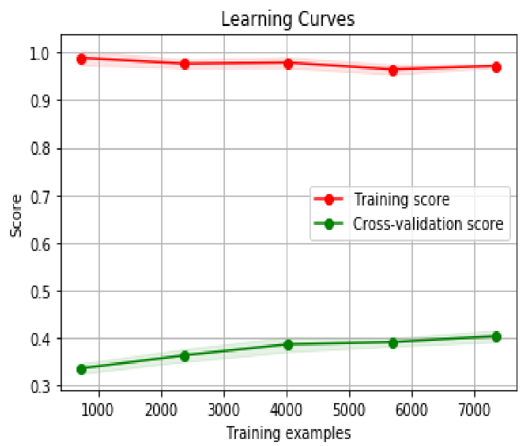

Figure 15: Learning Curve Artificial Neural Network

4.3 Preprocessing

We used two major types of features:

- Categorical Features - e.g., Gender, Type, Breed, Color etc.
- Textual Features - e.g., Name, Description.

The Categorical Features were preprocessed and converted to one-hot encoded features. The textual Features were passed through a Natural Language Processing pipeline, the final output of which was either 300 dimensional word embeddings or TF-IDF vector. In the Natural Langugage Processing pipeline, we first converted all words into lowercase. We then removed all non alpha-numeric characters and tokenized the description into token words. Finally, we removed the stop words like the, a etc. For converting the description token vector into 300 dimensional word embeddings, we used pre-trained word embeddings using skip-gram model on google news dataset from [8]. We found the individual word embeddings for each word in the sentence and then averaged these word embeddings to finally get the 300 dimensional embedding for Description feature. Similarly, for Tf-idf, we used Bag of words representation weighted by Term Frequency-Inverse Document Frequency for each word in the sentence. We then used singular value decomposition as a dimensionality reduction technique.

4.4 Ablation Study on Name Feature

We included Name textual feature in our intial experiments. We then perfomed an ablation study to see the importance of this feature and found that the feature was actually leading to a decrease in scores on the validation dataset. Noting this, we removed this feature from our further experiments. Thus, the observations in the Results section do not include Name feature.

5. RESULTS

This section is divided into 3 subsections. In the first section, we discuss the learning curves for each of the three models i.e. Support Vector Machine (SVM), Decision Trees and Artificial Neural Network (ANN). All the results mentioned in this sub-section are cross validation results using 5-fold cross validation. In the next section, we compare the performance of the 3 models on unseen test data. The evaluation metric is micro F1 Score for this performance comparison. The dataset used for this purpose contains Word2Vec embedding on the 'Description' attribute.

Model	Model Parameters	F1 Score on Test Data (Word2Vec)	F1 Score on Test Data (Tf-idf)
Support Vector Machine (SVM)	Kernel - Gaussian (RBF)	0.32	0.33
Decision Tree	Criterion - Gini Gain, Max Depth - 5, Minimum Number of Leaf Nodes - 10	0.37	0.35
Artificial Neural Network (ANN)	Number of Hidden Layers - 2, Neurons in 1st Hidden Layer - 1500, Neurons in 2nd Hidden Layer - 500, Learning Rate - 1e-5, Optimization Solver - 'Adam'	0.42	**0.44**

Figure 16: Performance Comparison between SVM, Decision Tree, ANN

Finally, in the last section, we perform an interesting study where instead of predicting among 5 output classes i.e. adoption speed from 0 to 4, we model the problem statement as a binary classification where 1 means that the pet was adopted in the first 100 days and 0 indicates that the pet was not adopted even after 100 days of the advertisement being listed on petfinder.my.

5.1 Learning Curves

We observe from the learning curves shown in Figure 13, Figure 14 and Figure 4.1 that in general with increase in sample size, the test accuracy is observed to be increased consistently among all 3 models i.e. SVM, decision tree and ANN. We also observe that Decision trees by using max depth of 5 and minimum number of leaf nodes 10 do not overfit since the test and training f1 scores are closer to each other especially with a sample size of more than 5000. We also observe that Neural Networks being highly non linear models give very high training scores. They also give the highest test f1 scores of around 0.42 with sample size of 7000.

We now discuss the learning curves for individual models below:

5.1.1 SVM

We observe from Figure 13 that the Test f1 score for SVM increases with increase in training data size with a slight decrease in training f1 score. The final training f1 score for SVM is around 0.38 and final test f1 score is 0.35. In the learning curve, the shaded area above and below the line indicates the standard deviation using 5-fold cross validation.

5.1.2 Decision Tree

For Decision Tree, we observe from Figure 14 that there is a much more consistent decrease in training f1 score when training data size is increased from around 1000 data points to around 7000 sample points. We also observe a parallel increase in test f1 score with the highest test f1 score observed at around 7000 data points of around 0.37.

5.1.3 Artificial Neural Network

We observe from Figure 4.1 that the Neural Network being a highly non-linear model gives very high training f1 scores on all sample training sizes. The test f1 scores increase with increase in sample size with the highest score of around 0.42.

5.2 Performance Comparison

The performance comparison is as given in Figure 16. We observe that ANN performs the best among the 3 models with f1 scores of 0.44 on test data. This score is observed when using TF-IDF representation for textual features. Our baseline model SVM gives f1 score of 0.32 using word embeddings and 0.33 when using TF-IDF. Decision Trees on the other hand perform better when using word embeddings i.e. the test score is 0.37. When using td-idf, the test f1 score for decision trees is 0.35.

5.3 Binary Classification

We observe from 17 that the highest model performance is by neural network model with a test f1 score of 0.78. This score is higher than the score of 0.44 which was the best in multi-class classification with 5 possible output classes. The binary classifiers can be more useful for shelters to know if the pet will be adopted within the first 100 days since it has a higher f1 score of 0.78.

6. CONCLUSION AND FINAL OUTCOMES

In this project, we used data mining algorithms to help shelters forecast the resources required for pet care. As

Model	Model Parameters	F1 Score on Test Data (Word2Vec)	F1 Score on Test Data (Tf-idf)
Support Vector Machine (SVM)	Kernel - Gaussian (RBF)	0.71	0.71
Decision Tree	Criterion - Gini Gain, Max Depth - 5, Minimum Number of Leaf Nodes - 10	0.69	0.7
Artificial Neural Network (ANN)	Number of Hidden Layers - 2, Neurons in 1st Hidden Layer - 1500, Neurons in 2nd Hidden Layer - 500, Learning Rate - 1e-5, Optimization Solver - 'Adam'	0.77	**0.78**

Figure 17: Binary Classification Performance Comparison between SVM, Decision Tree, ANN

promised in our project proposal, we applied 3 data mining algorithms SVM, Decision Tree and Artifical Neural Networks on the dataset from petfinder.my.

We used one-hot encoding for categorical features. For textual features, we used two encoding methods. In the first method, we used 300 dimensional word embeddings to encode the textual features. We also used TF-IDF encoding and present a comparison among the two. We followed the plan in our proposal and came up with interesting analysis which we show as visualizations in the Data Exploration section.

In addition, we performed an interesting study of modelling the given multi-class problem statement into a binary classification problem. We observed that the model performance improved significantly in the binary classification with highest scores reaching 0.78.

We hope this work is highlighted in the future and used by the community as a way to help shelters analyze and use the available resources efficiently to save millions of pets who are abandoned every year.

7. CONTRIBUTION OF MEMBERS

The execution of the different project stages were carried out in accordance with the plan created in the proposal document. The visualization was conceptualized by both authors and the most interesting graphs have been presented in the report. The preprocessing of the data and the algorithm implementation was performed by the authors using the pair programming paradigm in Python.

In preprocessing, we had to perform one-hot encoding, build a Natural Language Processing (NLP) pipeline, encode word embeddings and TF-IDF features. One-hot encoding and TF-IDF was performed by Nandita. The NLP pipeline and word embedding logic was written by Mohit.

The final report preparation was divided equally between the authors - Nandita focused on the first part of the report until the end of data exploration and Mohit worked on the Methodology and Algorithms section onward. Github was used for version control and collaboration. The repository is private and the rights are owned by the authors currently. The code can be made public on instructors' request.

8. REFERENCES

[1] Account from a shelter owner. https://akitarescue.rescuegroups.org/info/display?PageID=3247.

[2] Artificial neural networks as models of neural information processing. https://www.frontiersin.org/research-topics/4817/artificial-neural-networks-as-models-of-neural-information-processing.

[3] If everyone read this, the shelters would be empty. https://www.thedodo.com/dog-shelter-guide-adoptions-1532460278.html.

[4] Pet adoption basics: 5 myths about shelters. https://www.foundanimals.org/animal-shelters-pets/.

[5] Accuracy, precision, recall & f1 score: Interpretation of performance measures. https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-meas Nov. 2016.

[6] United states statistics of shelter intake. https://www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics, 2016.

[7] Pet features dataset. https://www.kaggle.com/c/petfinder-adoption-prediction/data, Feb. 2019.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 10 2013.

[9] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition, 2003.