Survival Analysis Project

Luciano Costa Marius Akre Stefanny Peraza Valery Zuñiga

View the project on GitHub

Introduction

For this analysis we wanted to work with a dataset different from health issues, and try to apply what we learned to other areas. In this case we took data related to animals in a shelter and their adoption[1]. It should be mentioned that the data frame does not contain the date of entry of the animals to the shelter, but the date of birth, so we had to adapt to the available data, since we really wanted to apply the analysis in another area, and therefore the results will be more a function of the age that an animal has before it is adopted, given that it is in the shelter.

In terms of survival analysis, the objective of this project is to determine the "risk" that pets in an animal shelter have of being adopted , based on different characteristics that are available in the chosen data set. It is worth mentioning that the data contains different outcomes for the animals: adoption, transfer, or no outcome. However, in our case we will only use a dichotomous output indicating whether it was adopted or not. The data ends on February 1st, 2018, after which date it is not known what happens to the animals.

We will also analyze how the following variables, when possible to include them, may influence the adoption or non-adoption of the animals.

Data dictionary:

- id: unique id for each animal
- age_upon_outcome: age of the animal when the outcome was determined
- animal_type: cat, dog, or … something else
- breed: breed of the animal
- color: color of the animal
- date_of_birth: date of birth of the animal
- datetime: date and time when the outcome was determined

[1] Data taken from Kaggle. We only use the `train.csv` data.

- name: name of the animal
- outcome_type: there are three possible outcomes: adoption, transfer, no outcome (euthanized, died)
- sex: sex of the animal
- spay_neuter: whether the animal was spayed or neutered: intact or fixed

Data preparation

EDA

At a quick glance at the data, we observe that in the case of categorical variables we will have to make modifications in order to be able to work with them, for example in the case of breed, color and name we have too many different types or levels, which would make it impossible to use them all. The variable name, on the other hand, has a significant number of null values.

Table 1: Data summary

Name	raw_data
Number of rows	54408
Number of columns	11
Column type frequency:	
character	8
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
age_upon_outcome	0	1.0	4	9	0	46	0
animal_type	0	1.0	3	9	0	5	0
breed	0	1.0	3	54	0	1812	0
color	0	1.0	3	27	0	475	0
name	16433	0.7	1	12	0	11826	0
outcome_type	0	1.0	8	10	0	3	0
sex	0	1.0	4	7	0	3	0
spay_neuter	0	1.0	5	7	0	3	0

In order to be able to work with this data, these will be the first modifications we will make:

- Since the variables breed and color have too many different combinations, we will remove breed from our data, and work a bit in a color categorization that allows us to use it.
- We have names for 69.8% of the animals, so in order to use it, we will create a dummy variable has_name, a variable that indicates that if there is a name or not.
- Also, we decided to treat the animal_type and sex variables as factors, so they can enter the model like that.
- For the duration data, we have the age_upon_outcome which is the difference between date_of_birth and datetime but this is not numeric, so we create our own variable time_to_outcome.
- We will work with spay_neuter as our grouping variable. We would like to verify if there are differences in adoption times if the animal is spay neutered or not. If we have an Unknown in the variable, we will remove the corresponding records as we do not know the information. Furthermore we do not have the sex of those records neither and most of them have early no outcome (i.e. early euthanized, died) in the range of time_to_outcome values. Therefore they may impact significantly our analyses while we do not have any information on their sex and spay_neuter to make reliable interpretation. Those Unknown records represent 8% of the dataset and we still keep almost 50 000 records after these removals.[2]
- We keep animal_type (Unknown sex and spay_neuter have been removed as described above). The only modification in this case will be that for animal_type, since there are only 5 observations in Livestock and it remains only 90 observations in Bird after the removals, we will join these categories to Other. In summary we have a majority of dogs and cats, where 48% are female and 52% male. In the case of spaying and neutering, we have 74% fixed and 26% classified intact.

animal_type	n	prop
Cat	18764	37.5
Dog	30669	61.4
Other	543	1.1

sex	n	prop
Female	23931	47.9
Male	26045	52.1

[2] The cross tables for the variables sex, animal_type, has_name, color_count, color_shade_intens, details on removed Unknown records, details on Kaplan-Meyer and Fleming-harrington estimators : see Appendix 1.

spay_neuter	n	prop
Fixed	37013	74.1
Intact	12963	25.9

In the case of the output variable, where the possibilities are adopted or not, we will create a new variable `outcome` for which we take `1` for adoption and `0` for no adoption.

outcome	n	prop
0	16797	33.6
1	33179	66.4

- Since we got some negative values for our variable `time_to_outcome` , which makes no sense as the `date_of_birth` should come before the outcome (unless some very rare or demanded type of pets get adopted before they are born), we will remove these five observations (5).

- Method to take into account the colors: We have too many colors but we expect this feature to have a significant impact on the choice to adopt or not. So, we build some subcategories in order to reduce the number of possible values and be able to take into account the embedded information. We try two kind of grouping : by number of colors and by intensity of shade.

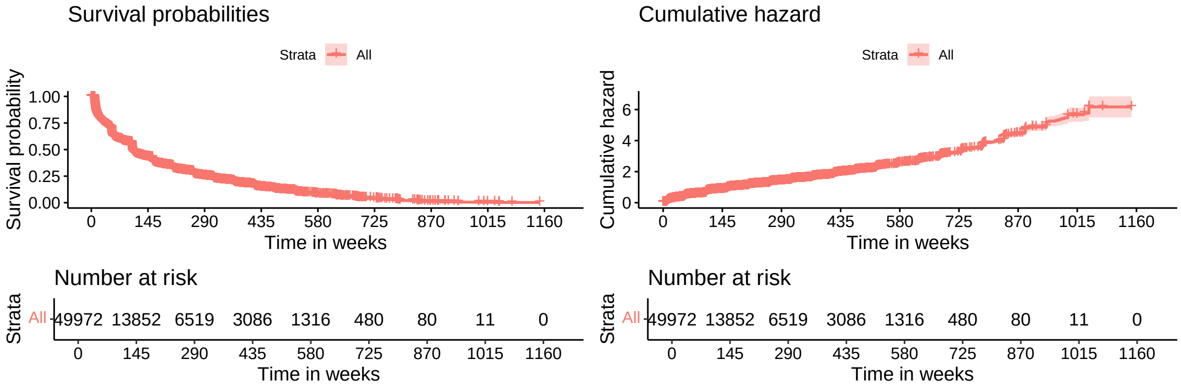
Now we chose the variable we will use as a clean data, check that they have the desired formats and do some previous analysis on the variables. The final total observation is 49,972.

Within the animals, there is not a difference that seems to be important according to sex, at least not proportionally. In the case of animals with names, more than 77% have been adopted against 34% in the case of those without names. Something similar happens with the animals that have been spayed or neutered, the proportion of adopted animals in this case exceeds 80%, while in the group of those that have not undergone surgery, the percentage of adopted animals is barely 26%. Lastly, dogs are not only the largest number of animals but also have the highest adoption rate, while among cats there is practically half and half between adopted and non-adopted, and among others type there is the lower adoption rate (36%). By looking at the distribution of the enriched dataset, it seems that having more number of colors increase the chance of adoption.[3]

[3] The cross tables for the variables `sex`, `animal_type`, `has_name`, `color_count`, `color_shade_intens`, details on removed `Unknown` records, details on Kaplan-Meyer and Fleming-harrington estimators : see Appendix 1.

Statistical Analysis

The next graph is the survival curve and the cumulative hazard function, for this data:



Nonparametric methods for censored data

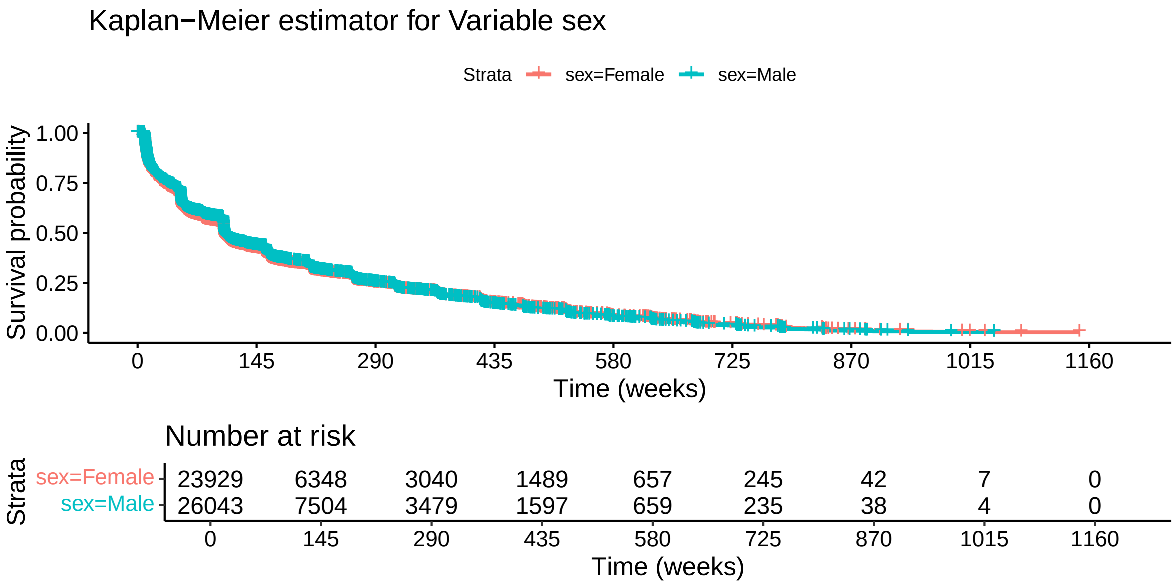
Using the Kaplan-Meyer estimator (KM) methodology we can see that the median survival time is 106 weeks (in this particular case the time from birth to adoption, and it's about 2 years). If we estimate the probability of not being adopted when the animal is one year old, we see that this value is 70.8% (for further information look at Appendix 2). We get similar results with Fleming—Harrington estimator. We will focus on the KM estimator in the following sections then.[4]

Nonparametric comparison of groups

Nonparametric comparison of 2 groups

Even if we do not have two groups in the dataset, for example with different treatments, we want to use this technique to analyze, for example, whether or not being part of the females group contributes to being adopted more quickly. So, the null hypothesis or question that we are going to try to answer is if the survival curves generated for the groups are the same, and the alternative hypothesis is that they are different. We will perform the logrank test.

[4] The cross tables for the variables `sex`, `animal_type`, `has_name`, `color_count`, `color_shade_intens`, details on removed `Unknown` records, details on Kaplan-Meyer and Fleming-harrington estimators : see Appendix 1.



```
Call: survfit(formula = Surv(time_to_outcome, outcome) ~ sex, data = data[,
    -1])

              n events median 0.95LCL 0.95UCL
sex=Female 23929  15767    105     105     106
sex=Male   26043  17410    106     106     107
```

```
Call:
survdiff(formula = Surv(time_to_outcome, outcome) ~ sex, data = data)

              N Observed Expected (O-E)^2/E (O-E)^2/V
sex=Female 23929    15767    15617      1.44      2.73
sex=Male   26043    17410    17560      1.28      2.73

 Chisq= 2.7  on 1 degrees of freedom, p= 0.1
```

As we can see both in the graph and in the logrank test, there is no statistical evidence to reject the null hypothesis of equality between Female and Male survival curves. This is consistent with our first basic analyses above on the cross tables. Hence, the sex does not make a statistical significant difference on the survival curves and the probabilities to be adopted.

In the `has_name` we get a difference between the curves and it looks like the adoption goes faster for those animal that don't have a name and are aged less than 2 years old at the moment of adoption. It is the over way round for the older pets. Surprisingly, the logrank test accepts H0 (p > 5% => cures are STATISTICALLY similar). This remind us that sometimes, the analysis should not be based only on the p-value (like some scientists thoughts nowadays). we can also look at other information that we have before communicating a final conclusion.

We can perform the same process with other categorical variables[5] : - `animal_type` : at least one of the 3 (crossing) survival curves looks different and H0 is rejected : at least one is statistically different. - `spay_neuter` : the 2 survival curves look different and H0 is rejected : they are statistically different. - `color_count` : the 3 survival curves look closed BUT H0 is rejected : AT LEAST 1 is STATISTICALLY different. - `color_shade_intens` : AT LEAST one of the 4 (crossing) survival curves looks different BUT H0 is accepted : they are STATISTICALLY similar.

Nonparametric comparison of more than 2 groups

We could also compare survival between 2 groups controlling for potentially confounding. For example, we could expect that `spay_neuter` feature may have an higher impact on the probability to be adopted for a Female as she is the one who carry their young.

By using stratification on `spay_neuter`, we can indeed see that for Intact `spay_neuter` groups the medians and the survival curves are less close from each other. Furthermore, we can see that the `sex=Female, strata(spay_neuter)=Intact` group has the highest probability (i.e. the lowest chance to be adopted) while it is the other way round within the "Fixed group" : Male are the lowest chance to be adopted. However, these results should be taken with cautious as the stratified logrank test display a p value = threshold = 5% (probably due to the fact that they are more pets in the Fixed group where it looks like there are less differences between Female and Male).[6]

Semi-parametric Cox regression

So far, we have consider non parametric models, i.e. without any assumption on the distribution of the data. Let's have a look on a semi-parametric model now (Cox model), where we will additionally assume some distributions on the covariates but without any type of assumption on the component ratio of hazard.

```
Call:
coxph(formula = Surv(time_to_outcome, outcome) ~ animal_type +
    has_name + sex + spay_neuter + color_count + color_shade_intens,
    data = data)

  n= 49972, number of events= 33177
```

[5] for further information look at Appendix 2
[6] for further information look at Appendix 2

```
                            coef exp(coef) se(coef)       z Pr(>|z|)    
animal_typeCat           0.23545   1.26548  0.07279   3.235 0.001217 ** 
animal_typeDog          -0.01486   0.98525  0.07232  -0.205 0.837231    
has_nameTRUE            -0.10005   0.90479  0.01749  -5.720 1.07e-08 ***
sexMale                 -0.01872   0.98145  0.01110  -1.688 0.091487 .  
spay_neuterIntact       -0.66741   0.51303  0.01899 -35.153  < 2e-16 ***
color_count2             0.10392   1.10951  0.01488   6.985 2.84e-12 ***
color_count3             0.02813   1.02853  0.03509   0.802 0.422733    
color_shade_intensDark   0.12966   1.13844  0.03378   3.839 0.000124 ***
color_shade_intensLight  0.09097   1.09524  0.03374   2.696 0.007015 ** 
color_shade_intensMedium 0.11213   1.11866  0.03271   3.428 0.000607 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                        exp(coef) exp(-coef) lower .95 upper .95
animal_typeCat             1.2655     0.7902    1.0972    1.4595
animal_typeDog             0.9853     1.0150    0.8550    1.1353
has_nameTRUE               0.9048     1.1052    0.8743    0.9363
sexMale                    0.9814     1.0189    0.9603    1.0030
spay_neuterIntact          0.5130     1.9492    0.4943    0.5325
color_count2               1.1095     0.9013    1.0776    1.1423
color_count3               1.0285     0.9723    0.9602    1.1018
color_shade_intensDark     1.1384     0.8784    1.0655    1.2164
color_shade_intensLight    1.0952     0.9130    1.0251    1.1701
color_shade_intensMedium   1.1187     0.8939    1.0492    1.1927

Concordance= 0.615  (se = 0.002 )
Likelihood ratio test= 1936  on 10 df,   p=<2e-16
Wald test            = 1707  on 10 df,   p=<2e-16
Score (logrank) test = 1752  on 10 df,   p=<2e-16
```

The Cox model shows that 3 variables are not statistically significant (animal_typeDog, sexMale, color_count3), meaning those categories are not significant different from the base category.

About the colors, known shade intensity makes statistically a significant difference to be adopted faster comparing to pets having an Unknown intensity (~ 10% more chances). Having 1 or 3 colors does not seem to make a difference but, having 2 colors comparing to 1, would increase the chances to be adopted by +10% to +12%. Having 2 colors makes statistically a significant difference comparing to having 1 color, by increasing the chance by 10% to be adopted.

Automatic model selection based on AIC

However, using all these covariates may not be the best choice to get the best model (from the AIC criteria point of view for example). By applying an automatic model selection based on AIC we get the conclusion that we should keep all the covariates in addition to an intercept, to have the best model (highest AIC).[7]

Prediction

A model can be used in two main ways : to analyze or study interactions between the covariates like we did or to predict. Let's see our predicted survival proportion for the whole data. We can see that the estimated probability of not being adopted within 1 year is slightly higher (71.4%) than the one of KM model (70.8%) but they remain close.[8]

Now, we can also do a prediction and verify how this survival changes depending on the `spay_neuter` variable for example.

As expected, animal which were spay neutered (Fixed), have shorter survival time, that means they were adopted faster.

Model diagnostics

After getting these model, we should verify the assumptions of these model. Let's do it for the Cox model for example.

Martingale residuals and Proportionality of hazards

The residuals are not so well formed (not around 0). There are many outliers. Schoenfeld residuals are far from zero and not in the boundaries. Furthermore, the plot of proportionality of the hazards does not look good : H0 (beta = beta(t)) is rejected. We can try stratification or truncate to fix that. However, stratification did not make the assumption validation better[9].

The model would not be validated from a theoretical point of view, although the results of our previous analyses of the models seem very consistent to the behaviors that we could have expected a priori. Indeed, conclusions of the analyses seem also consistent to what we may observe in reality for pet adoption. We could conclude by taking all previous results with cautious and also remind this common aphorism: "All models are wrong, but some are useful", George E. P. Box

[7] for further information look at Appendix 3

[8] for further information look at Appendix 3.

[9] for further information look at Appendix 4