Kaggle competition report:
Animal Shelter outcome prediction

(Bulat Nasrullin, Ildar Nurgaliev, Semen Zorin)

Data description:

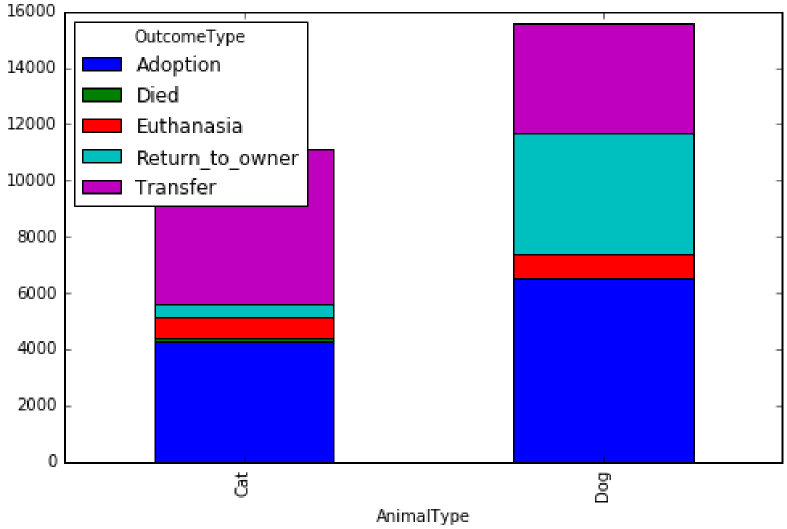
Link of the competition: https://www.kaggle.com/c/shelter-animal-outcomes
The aim of the competition is to help shelters' personal understand trends in animal outcomes. These insights could help shelters focus their energy on specific animals who need a little extra help finding a new home.

Team name: **IDST_mini**
Features:

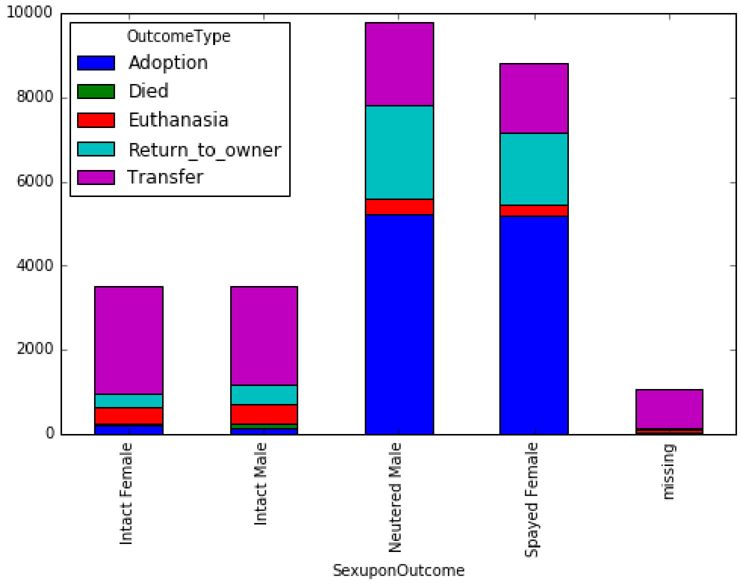
AnimalID	All animals receive a unique Animal ID during intake
Name	A name or anonym
DateTime	Acceptance time
OutcomeType	Adoption, Died, Euthanasia, Return to owner, Transfer
OutcomeSubtype	Additional info about outcome type, not presented in test data
AnimalType	Dog or cat
SexuponOutcome	Male and female combined with sex ability
AgeuponOutcome	From some days up to years (time in shelter passed)
Breed	Enormous amount of different breed and their mix
Color	Different colors and their mix

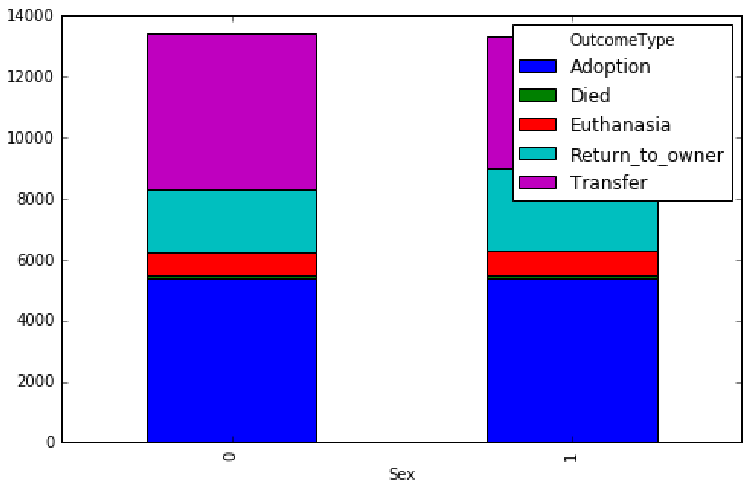
Data analysis and feature reduction

At the first step of data analysis we conduct visual stacked bar plots for understanding the similarity of outcomes with respect to a feature. In the plot below we see that the relative outcome of a dog and of a cat is the same except for the *Return_to_owner* outcome which is significantly higher in case of dogs.

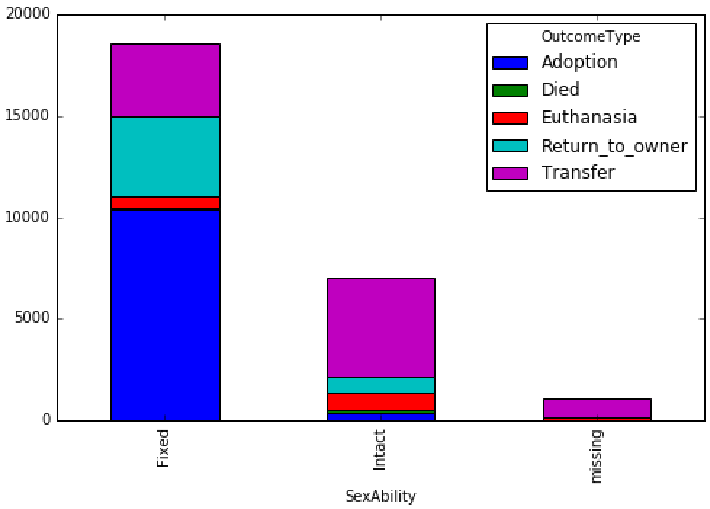


At the next step we decided to reduce features value to achieve more expressive feature representation. The next feature to consider was ‘SexuponOutcome’. Original outcome types based on SexuponOutcome field and reduced Sex field may be seen on the figure below:

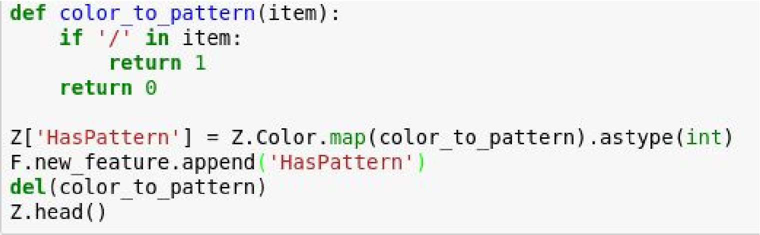




It became clear that there is no difference just in sex after sex elicitation as a feature, the difference of outcomes is represented in fixed and non fixed sex ability, thus we reduced combinations to Fixed, intact and missing, their outcomes are presented on the right plot (**SexAbility**).



As the next feature we decided to reduce **color** feature from its different combination. As analysis show, cats have 165 different values of a color and dogs have 295. The reduction is just simple splitting of mix color to it real color and one hot representation of it occurrence of an animal. Before reduction, totally we had 411 different color, after reduction it become 58. Also we decided to elicit **pattern occurrence**:

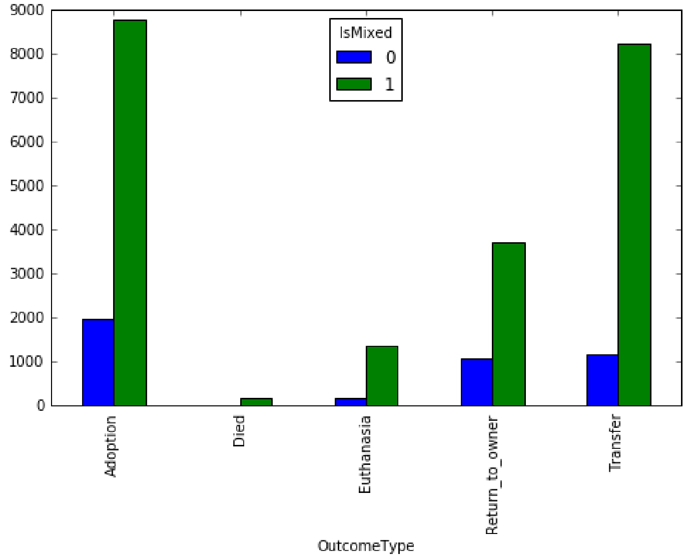


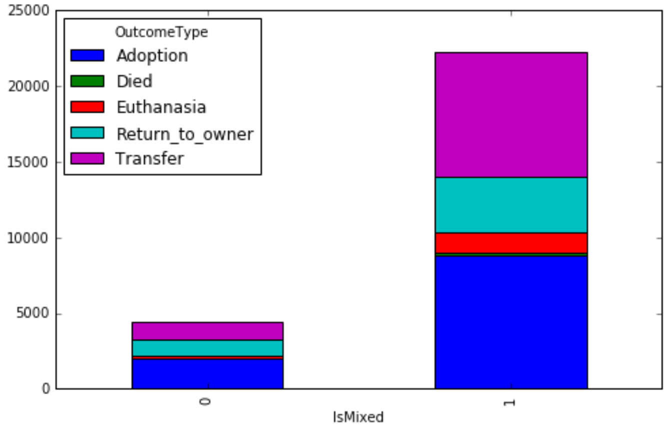
```
def color_to_pattern(item):
    if '/' in item:
        return 1
    return 0

Z['HasPattern'] = Z.Color.map(color_to_pattern).astype(int)
F.new_feature.append('HasPattern')
del(color_to_pattern)
Z.head()
```

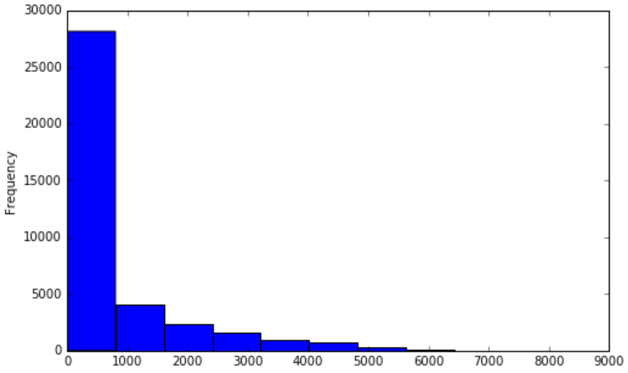
This new feature had good affection to prediction. It could be described as people do came after the animal because of a particular color but it interesting pattern presence.

As in previous experiment, we generated feature of pattern occurrence in color and it had good affection, so we considered it could play the same good role in breeds, namely saying if we could elicit mixed breed as nex feature, but results show us there is the same distribution among mixed and unmixed (plot below shows).

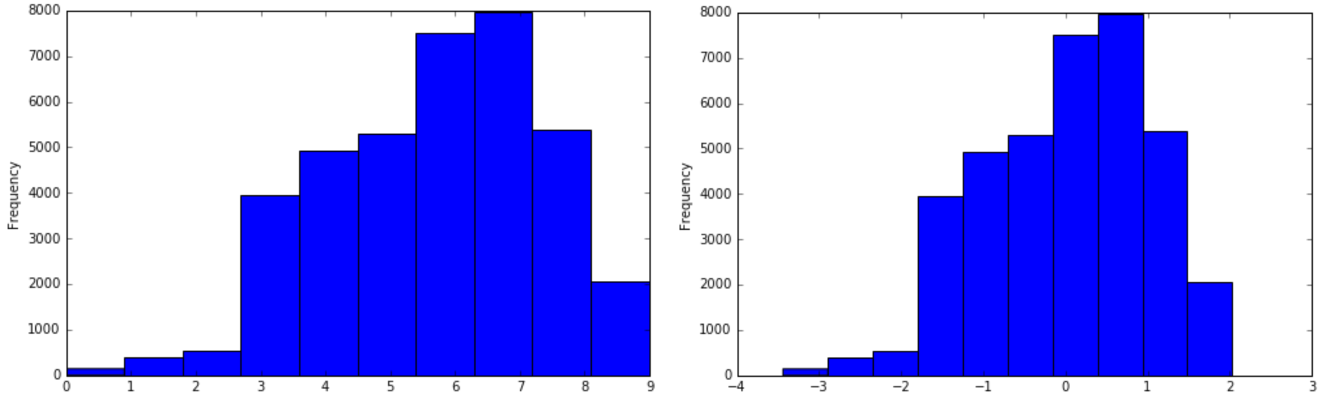




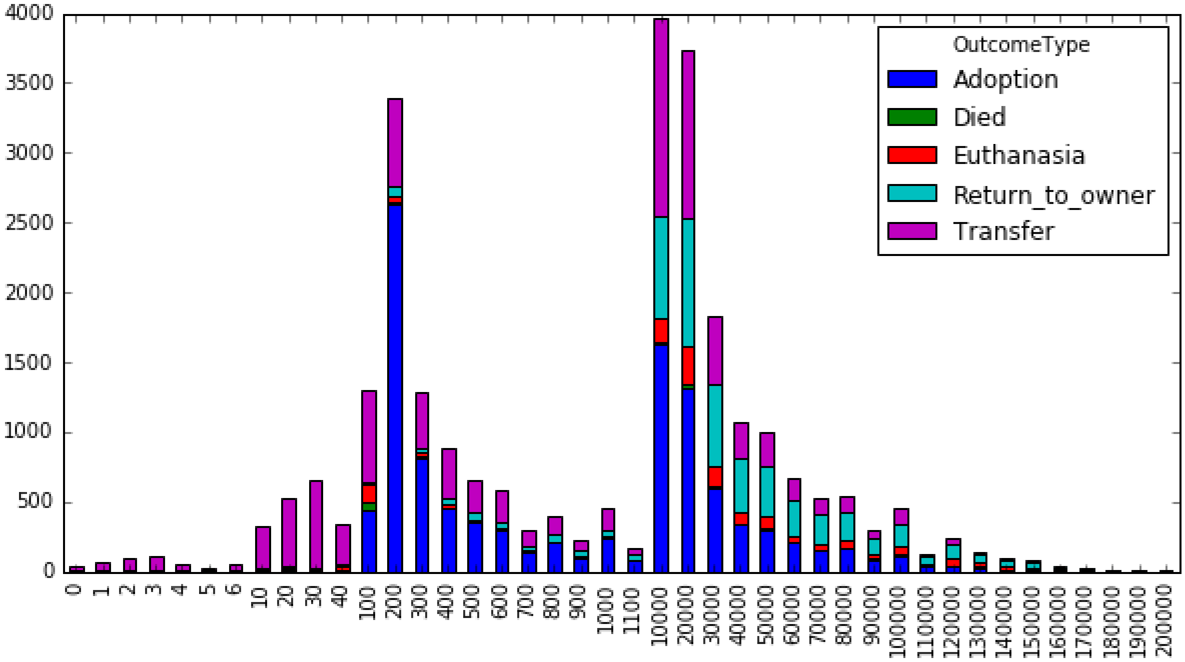
Finally we just reduced breeds of dogs. Because cats have just 70 different breed while dogs have 1608 that seems to be a very high number.



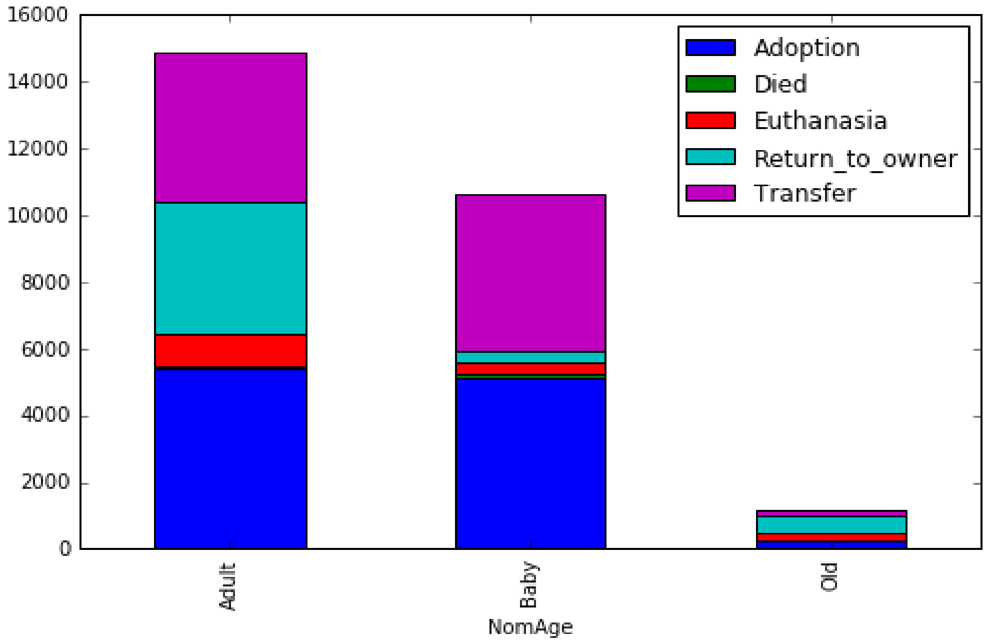
Dogs were combined to breed groups using information from Wikipedia. There was also a very good feature AgeuponOutcome, but as it was represented as a string date, we transformed it to days count and plotted its distribution among animals (histogram). The histogram shows us right curved distribution, we could transform it to bell curved distribution by log scaling. The histogram below shows log transformation of days and its z-score preprocessing on the right plot.



In the plotting of Outcome distribution with respect to days, we got that Adoption and Transfer are more centered in distribution, while *Return_to_owner* is mostly right skewed.

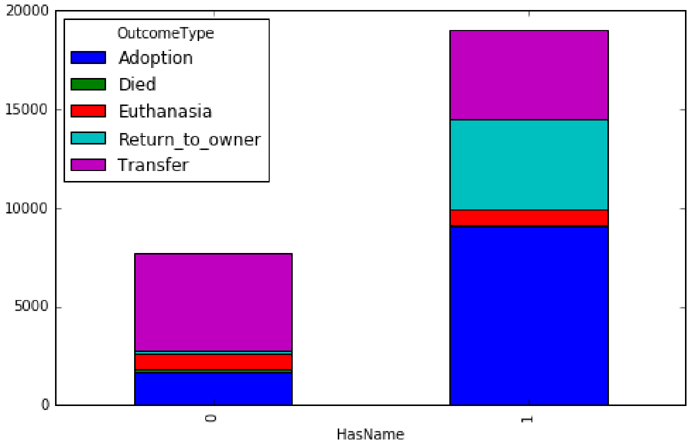


As we see on the histogram plot, most of the animal are younger than 3 years. We could generate features like young, middle, old. Because from the plot above it is clear that old dogs more likely to be returned back to owner (right skewed distribution of *Return_to_owner* with respect to DaysNorm).

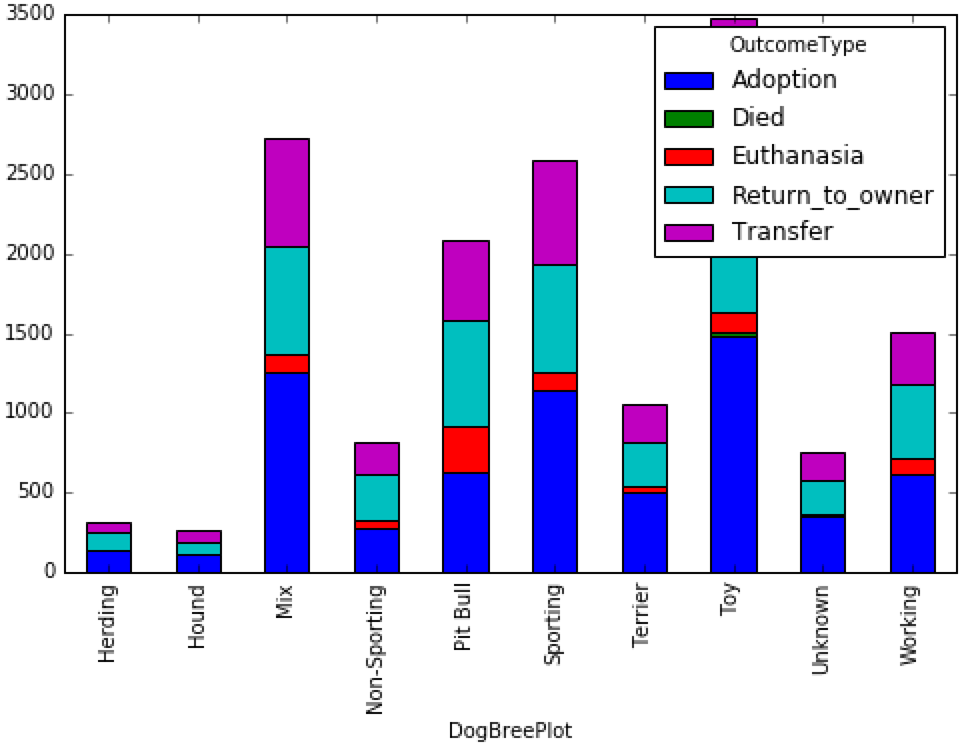


As it is expected baby cats and dogs have much higher chances to be adopted or transferred than to be returned to owner, while older animals more likely to be returned, while adult has more chance to be euthanasia owner (age_baby = 7 * 31 , age_young = 365 * 3 , age_adult = 365 * 10).

The next feature generated was name occurrence “HasName”. It is hard to play with a particular name of an animal while it became clear that animal with a name is more suspected to be adopted or returned to an owner.



The last feature to analyse and reduce was breed. We found out that cats have just different breeds while dogs have 1608. Thus we decided to leave out cats breed as is, while dogs breed mostly repeated, so we considered this is very redundant aspect to be reduced by grouping dogs to 'Herding', 'Hound', 'Mix', 'Non-Sporting', 'Pit Bull', 'Sporting', 'Terrier', 'Toy', 'Unknown', 'Working'. We found some stop words in breed nomination of a dog, as for example ‘shorthair’, ‘longhair’ or ‘Rough’ etc. As a result we reduced dog breeds(1608 modified types of breeds mostly same) up to 10 dog groups. The Outcome distribution is presented below.



Chosen features

One-hot-encoded AnimalType	Dogs and cats
One-hot-encoded SexAbility	SexAbility is reduced SexuponOutcome feature
One-hot-encoded ColorOne	ColorOne is feature generated after color values reduction
DaysNorm	Representation of AgeuponOutcome text date as days count that was preprocessed by log transformation and z-scaler.
HasName	Name occurrence
HasPattern	Pattern occurrence
Baby	Age is lower than 7 months

Adult	Adult and young animals
Old	Over 10 years
One-hot-encoded CatBreed	Cats have just 70 breeds
One-hot-encoded dog groups	Dog breed was reduced to dog group (10)

Selected model

The train data was splited randomly to train and test data for model evaluation and set-up a particular model. Train data rows count = 26729, test rows count = 11456. Afterwards there was generated KFold for finding out best features of a model by GridSearch.

To solve the initial problem, several classification models were tried with different parameters.Below one can see a table with tested classifiers and their results on public dataset.

Classifier	Public score
Random Forest	1.95798
Multinomial NB	1.42597
AdaBoost	1.56860
SVC with tf-idf encoded breed and color	1.18855
SVC	0.92239

The first one was to use multinomial Naive Bayes classifier, as it is frequently “suitable for classification with discrete features”. As it turned out in future, its results were not so good in comparing with other classifiers. According to statistics, most of the competitions are won when ensemble methods are used. Two of them were tried to predict classes probabilities: AdaBoost and Random Forest. Both of them slightly improved the results of NB classifier, though, not having the best results among tried classifiers. SVC classifier gave us best predictions in both main ideas of feature processing (described later).

Model evaluation

Two major attempts were made during working with the features. First one was just simple grouping objects based on some criteria and using one-hot-encoding for each of the feature. It showed better results then the second method, where *Breed* and *Color* features were processed using tf-idf vectorizer. The underlying hypothesis for this method was that some breeds or color combine better, getting the resulting object (i.e. dog or cat) more interesting and likely to be adopted. For this to be done properly, it was necessary to split *Breed* and *Color* values into separate words. After closer look on the breed, one can see that “Mix” word is frequently used to represent a pet with parents of two different breed. In this case, we need to take into account three-grams (for both parents breeds and mix word itself). As for color, three-colored pet are hardly ever represented in the set, one could take only 2-grams. Nonetheless, hypothesis did not produce good results, so the model with all the one-hot-encoded features was taken as main one.