Project „Titanic”

Your task is to build a predictive model based on real dataset – the list of "Titanic" ship passengers. This model, based on variables selected from 13 predictive attributes, will allow you to determine the value of the outcome variable ("survived"), i.e. the attribute the model is going to predict. It tells us whether a given passenger survived the disaster or not and takes values [0, 1], where 0 = No, 1 = Yes. The full dataset, including variables outside the list below, is available at https://github.com/jbryer/CompStats/blob/master/Data/titanic3.csv.

To determine the value of "survived" label, we will use subsets of 10 explanatory variables:

- pclass – cabin class,
- name – passenger’s name (including title),
- sex – passenger's gender,
- age – passenger's age,
- sibsp – number of spouses or siblings on board,
- parch – number of parents or children on board,
- ticket – ticket number,
- fare – ticket fare,
- cabin – cabin number,
- embarked – port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

The other features ("boat", "body", and "home.dest") should be dropped, since they are either irrelevant or may lead to data leakage.

At the beginning, you should make some assumptions regarding the model, for example, when to end the "play" with its improvement. Let's take the usual level of accuracy achieved in such analyses as a reference point, around 80%. A result no worse than this number can be considered a success and you can finish working on the project – unless your ambition has not been satisfied. Remember to evaluate the model on test data, never on training data!

1. Data set analysis

Start by downloading the data, then load it into Dataiku DSS. To take a closer look at the data, perform exploratory data analysis. This will allow you to view basic metrics for each column, such as minimum values, maximum values, averages, medians, distributions, correlations, etc.

Try to find errors in the data – missing values, incorrect interpretation of variable types, etc. Consider which attributes to choose for analysis (in terms of their significance for model predictions, potential data leakage problems, etc.), whether and how to fill in missing data, perform transformations, etc.

2. Feature engineering

For example, regarding filling in the values of the "age" column, after analysing the data, you may come to several conclusions, which are e.g. as follows:

- All unmarried women with the title "Miss" who have no children are ladies with an average age of ... years.
- Gentlemen with the title "Master" are bachelors with an average age of about ... years.
- "Sir", "Mr", "Ms", and "Mrs" denote mature individuals with a few exceptions.
- All gentlemen with the title "Dr" are mature men with an average age of ... years.

Fill in the age for all of the above groups using the average for the respective group. Similarly, for all other people without a specified age, fill in their age with the average for their gender.

Create three additional variables that you will use later in the experiment:

- "family.size" – by adding together two existing values: "parch" and "sibsp", plus the number 1 corresponding to the passenger itself. Intuitively, one can assume that family size may have had a significant impact on whether someone survived or not.
- "age.range" – this is a categorical variable that assigns a passenger to one of four age categories: "baby", "child", "teenager", or "adult". Here we have in mind the principle of "women and children first". Assume ages 6, 12, and 18 as the division points.
- "mpc" – a variable that is intended to "highlight" the chances of survival for children and first-class passengers. It is the result of multiplying a person's age by the class in which they travelled. For example, a 5-year-old child traveling in first class (mpc = 5) had much better chances of surviving the disaster than a 70-year-old man traveling in third class (mpc = 210).

These three additional variables will be a good starting point for achieving the highest possible score in your machine learning process.

3. Metadata editing

Before filling in missing data, you should check (and possibly change) the types of individual columns. Pay particular attention to the columns: "survived", "pclass", "embarked", and "sex" – they should represent categorical data. What type of values should the "fare" column contain?

4. Missing data completion

Now you need to fill in missing values for the remaining attributes. You can use one of the imputation methods offered by Dataiku DSS – use the "Prepare" recipe and appropriate formulas. Try to do this intelligently, using the knowledge gained during exploratory data analysis – for example, using the mean values not for the entire dataset, but for individual subgroups of records.

5. Trimming outliers

Start by analysing the values of individual attributes for outliers. The best tool for this is a scatter plot, where we plot the same attribute on both axes, or a box plot. This makes it clear which values are outliers. For example, in the visualization of the ages of individual passengers, we may notice that for the "age" column, we should remove all values above 67 and replace them with... well, what? We basically have three options – we can use the previously chosen boundary value (67 in our case), the average value, or mark them as

missing data. Consider which approach would be most appropriate here. Similarly, the ticket price paid by individual passengers should be treated. NOTE: Trimming outliers should always be done before data normalization.

6. Normalization of numerical data

All numerical data must be normalized/standardized. Why? Well, during machine learning, variables with larger values may be perceived as more important by the algorithm. To better explain this, let's use an example from the Titanic passenger data. A passenger named "Sandstrom, Miss. Marguerite Rut" fortunately survived the Titanic disaster. The data clearly shows that she paid a fare of 16.7 and was only 4 years old at the time of the journey. The algorithm may conclude that the fare had a significantly greater influence on her survival, which may not necessarily be true. Almost everyone (especially after watching James Cameron's "Titanic") can hypothesize that women and children had the greatest chance of survival... That's why normalization must be done.

Of course, normalization should be done for numerical variables (we don't want the algorithm to normalize categorical data). Choose "MinMax" as the method used for data transformation.

7. Selection of learning algorithm

Without a doubt, we are dealing with a binary classification problem here. Based on the given attributes, we need to determine one of two values: [0, 1], [Yes, No], [True, False], etc. In our case, the target variable we want to predict is the "survived" column.

Finally, you should choose a maximum of 6–8 attributes that will serve as predictors. From experience, it is known that with such a number of attributes, decision trees or random forests perform best for classification problems. However, don't end your search for the "perfect" algorithm there – try others available in Dataiku DSS as well.

To compare the results provided by different algorithms, you need to divide the data into training and validation sets. Randomly divide the data, with 80% being training data. The next step is to choose the predictive variables – try starting with the columns "sex", "survived", "age", "age.range", "pclass", and "fare".

8. Model evaluation, tuning, and analysis

Now you need to train the selected models (algorithms) on training data (use the test subset to check for overfitting), evaluate their effectiveness, and compare the results obtained – finally, further attempts to improve them through continued feature engineering and manipulation of hyperparameters. Also, examine the developed models in terms of explainability and conduct analyses of how individual explanatory variables affect the target variable – perform these four steps analogously to the "Machine Learning Basics" course (https://academy.dataiku.com/machine-learning-basics) at Dataiku Academy.

Have fun!