Assignment #1
Biostat 202

Introduction to Orange Data Mining (due 8/6/2020 at 11:55 PM)

Description: The purpose of this lab is to familiarize you working with "widgets" and building "workflows" in Orange Data Mining. In Orange, each step in the data analysis is represented by a widget and the sequence of actions represented as a series of linked widgets is called a workflow. In this lab we will build a workflow to analyze passenger survival following the sinking of the RMS Titanic. To do this, we must read data into the software, define their roles, do a data quality check for missing data, transform some of the variables to facilitate analysis, generate simple tabular descriptive statistics, build some predictive models, and check their performance by dividing our dataset into a "training" and "test" subset.

Skills:

- Link widgets together to form workflows in Orange
- Import data
- Perform basic data quality check, missingness and skew
- Transform variables
- Drop variables/values
- Calculate descriptive statistics
- Train predictive models and evaluate performance

Steps: *Please follow the steps below and answer questions marked in* ***bold in a separate document****. Include a screenshot of your final workflow.*

For this lab we will use the data on survival of the passengers from the famous sinking of the cruise ship, RMS Titanic. The dataset is available on the course CLE website under Lab 1 ("titanic_survival.csv"). It contains the following variables:

PassengerID: passenger ID variable
Survived: 1=survived, 0=died
Pclass: passenger cabin class, 1=First, 2=Second, 3=Third
Name: Name of passenger
Age: Age of the passenger
Sibsp: Number of siblings/spouses on board
Parch: Number of parents/children on board
Ticket: Ticket number
Fare: Passenger fare
Cabin: Cabin passenger was assigned to
Embark: Port in which passenger boarded the Titanic (C = Cherbourgm, Q = Queenstown, S = Southampton).

Our goal will be to build a model to accurately predict who survived and who did not. You will make use of the Orange Data Mining software, installation instructions for which are available on the course CLE website under Lecture 1.

1. Because Orange depends on a graphical interface to select widgets, connect widgets via channels (links), and build workflows, it is important to get an idea of how a workflow is built up by watching its construction. The basic mechanics of Orange can be reviewed via the Orange YouTube channel: https://www.youtube.com/c/OrangeDataMining/videos. Students are expected to complete assignments on their own or in groups but today will feature more hands on instruction to make sure everyone is comfortable with the basic functionalities.

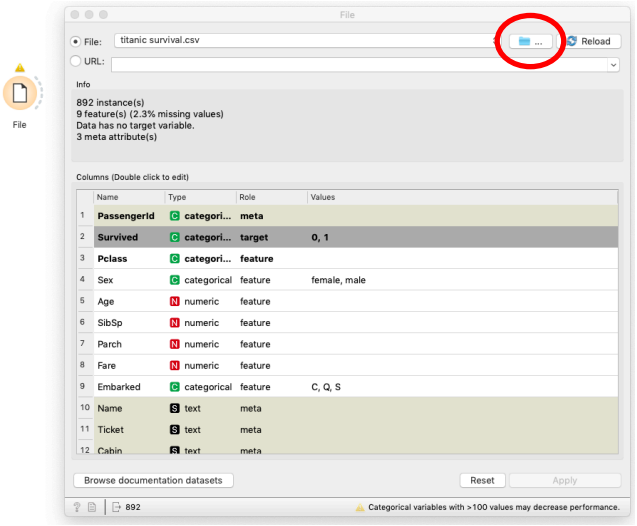
2. Next download the dataset "titanic_survival.csv" from the course website. Open it with Excel or a spreadsheet program to get a sense of the data. The extension ".csv" is short for comma separated values, so named because if you look at the file in its raw format (e.g., using Word) the data values are separated by commas.

3. Now we will start building a workflow. A workflow represents the sequence of operations from the data source to the final model or output. Start Orange and click on "New" under the File tab.

4. Along the left column, you will find five categories of widgets: Data, Visualize, Model, Evaluate, and Unsupervised. The five categories are described below.
    a. Data: widgets for importing and manipulating data
    b. Visualize: widgets for visualization of data
    c. Model: widgets for modeling the data, specifically to predict an outcome
    d. Evaluate: widgets for evaluating predictive performance
    e. Unsupervised: widgets for modeling the data, specifically to uncover structures or patterns

5. Click on the Data tab and click on the "File" widget. A widget will appear in your main window titled "File". Right-clicking on the widget will allow you to perform actions such as open, rename, remove, duplicate, copy, and most importantly, access help which provides a definition of the widget and examples.

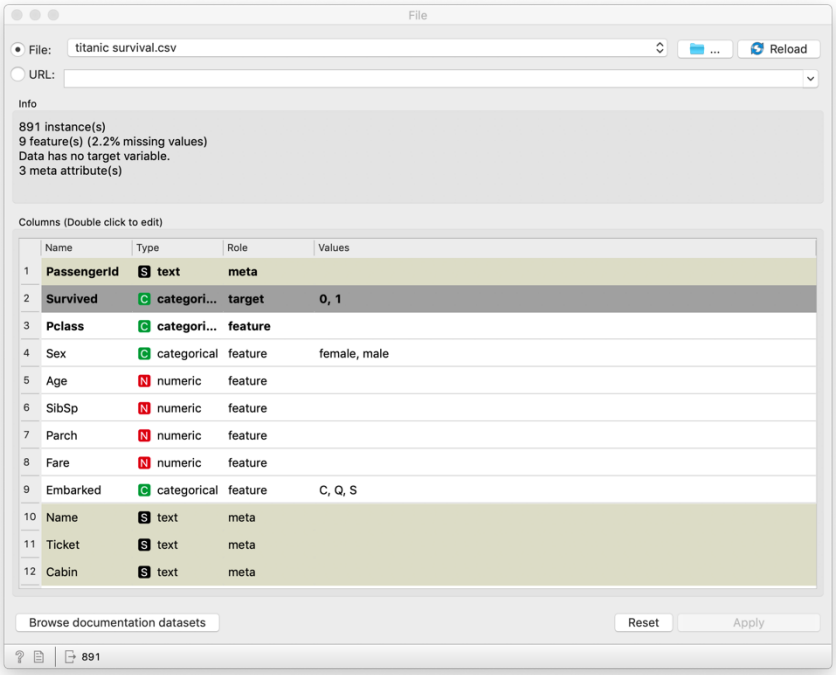
6. Double-click on the widget to open it up. Your screen should look something like the figure below. Click on the folder icon (circled in red) and find the "titanic_survival.csv" you downloaded. Click on the Open button to import the file into Orange.



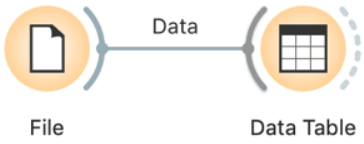
7. Once the data is imported, we can see some summary information. There are 891 instances (records/rows, i.e. passengers), 2.2% of values are missing, etc… This information is useful in order to broadly describe the structure of the data.

8. Next we need to tell the software what the type and role of the variables will be in the analysis. Orange will read the data and try to figure out this information automatically. With the "File" widget still open, you can see the default variable types and roles assigned by Orange. The types are as follows: numeric is a continuous variable (e.g. weight), categorical is a discrete categorical variable (e.g. race/ethnicity), text is an unstructured sequence of words (like this assignment), and datetime is a formatted date and time (e.g. 2016, 2016-12-27, 2016-12-27 14:20:51). The roles describe the function of the variable in our workflow: feature is a predictor or independent variable, target is an outcome or dependent variable, meta variable which provides information that is not used in analysis, and skip means the variable will not be displayed in subsequent steps. Values for only categorical variables are displayed.
   Q1: Which variables do you think should be used to predict survival?

9. We wish to predict survival, so we change the role of the *Survived* variable to target. This is accomplished by double-clicking the Role value associated with *Survived* and selected target from the dropdown. For this first exercise we will just use a subset of the variables as features for prediction. Orange has already designated *Name*, *Cabin*, and *Ticket* as Type meta. Change *PassengerID* type to Type text and role to Role meta since it uniquely identifies the passenger but is not useful for prediction. Change *Pclass* to Type categorical (can you guess why it thought it was Type numerical?). Variables designated as Type feature will be used in the model but those designated as Type meta will not be used. Click Apply and exit the widget. Your "File" widget should now look like the figure below.



10. Click on the Data tab and select a “Data Table” widget which will now appear in your workflow. Connect the “File” widget to the “Data Table” widget by selecting the grey half semi-circle on the “File” widget, holding your cursor down, and dragging a line to the “Data Table” widget. The channel (link) should look like this:



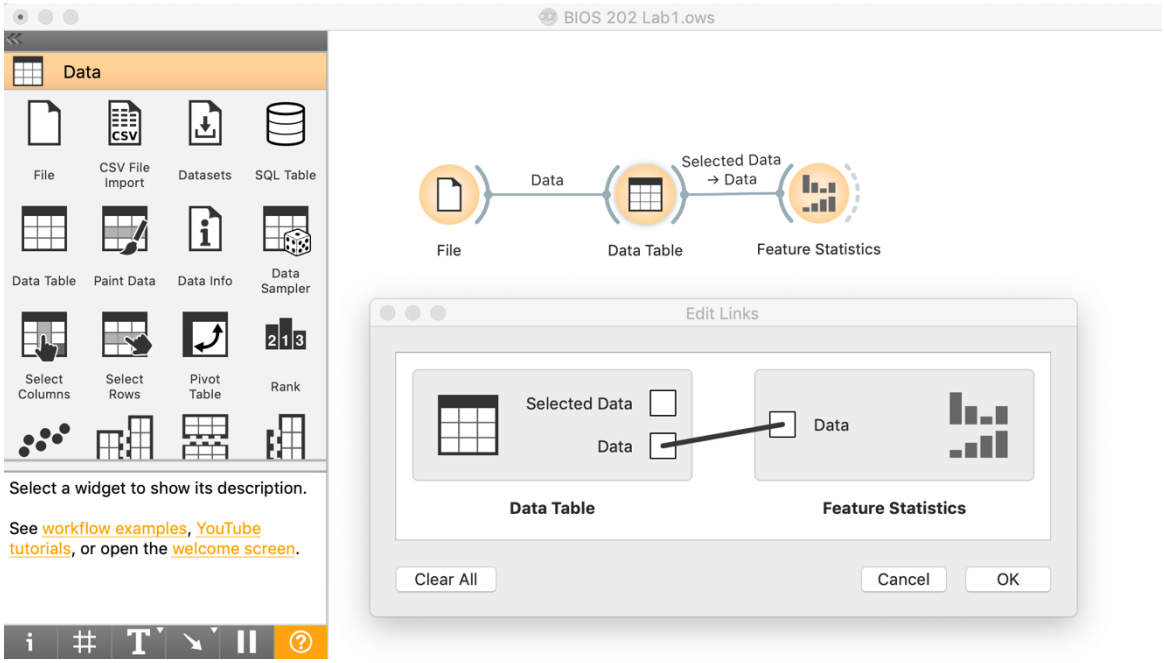
11. Note on connecting widgets via channels in a workflow. By linking the grey half semi-circles of widgets with channels, you provide a link through which information flows. This can include data, or even output from a widget. You can create channels (links) between widgets two ways:
    a. Select the widget from the left column and repeat the procedure described in 10 to connect your existing widgets to the previous widget.
    b. Alternatively, you can begin by selecting the origin widget and hold your cursor down and drag it to an empty space in the workplace. When you release your cursor, a menu of widgets to choose from will pop up.

12. Open the “Data Table” widget. You should see the imported data, with passengers organized by row and variables by column. Variables with Role target, meta, and categorical are denoted with grey, green, and white shading, respectively. Experiment with the Variable checkbox options. You will notice that many values for *Cabin* variable and some for *Age* variable have ? symbols. This

means the data is missing (confirm with original “titanic_survival.csv” file). We will discuss how to deal with missing values later. The “Data Table” widget is crucial for inspecting your data and verifying the quality of import and subsequent transformations. You will use it frequently throughout the workflow. Close the widget.

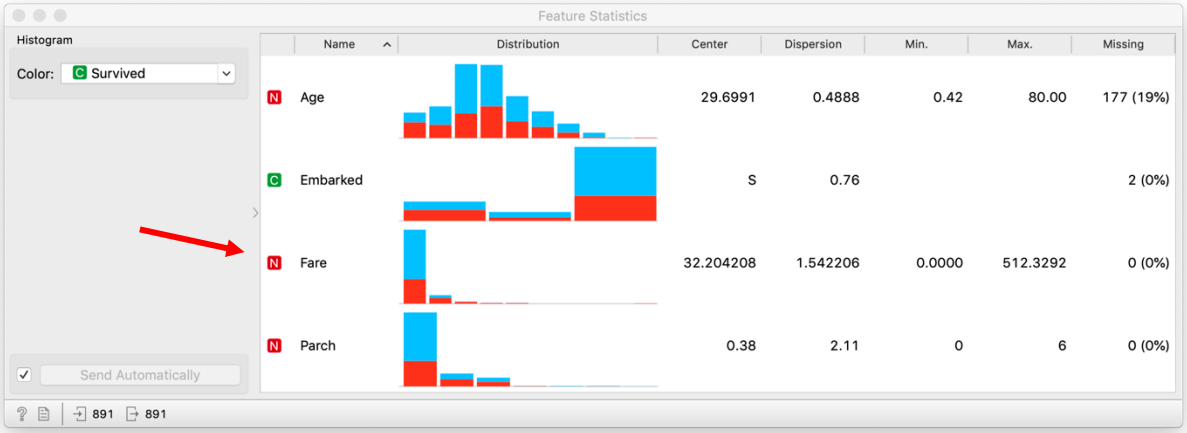
13. We will now do some exploratory data analysis and quality control checks of the data. From the Data tab, connect the “Data Table” widget to a “Feature Statistics” widget.

14. Note on editing channels in a workflow. Channels control the flow of information through a workflow and are as fundamental as the widgets themselves. For example, double-click the newly created channel between the “Data Table” widget and the “Feature Statistics widget”. A menu pops up that shows how data flows, in this case Selected Data or Data flows from the “Data Table” widget to the Data for the “Feature Statistics” widget. Whatever information is selected to flow down the workflow will be used. In this case, select Data to flow from the “Data Table”. Click OK to confirm. It is always worth checking what the channel properties are between widgets when you create them, this can be a challenging and confusing aspect. Your workflow and most recently created channel should look like this:



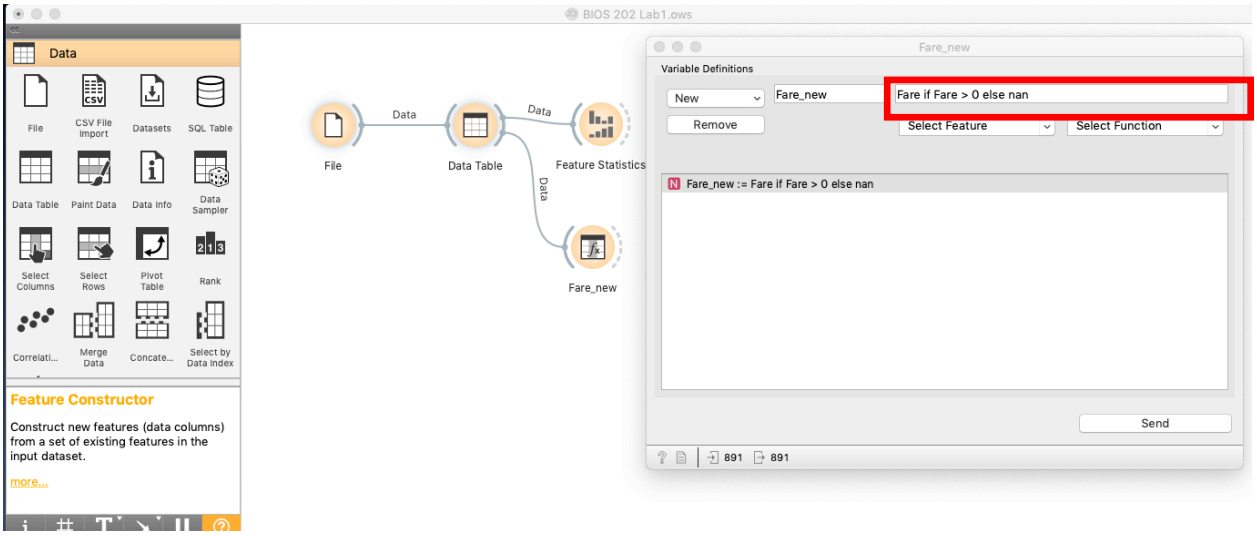
15. Open the “Feature Statistics” widget. Orange will display each of the variables in a histogram (for numerical variables) or a bar chart (for categorical variables) with color coding for survival or not (since we have designated *Survival* as the target variable). If you wish to stratify by another variable, select it from the

dropdown menu for Color (e.g. *Sex*). Note that *Fare* highly skewed to the right, which we will consider below. **Q2. Looking at the figure, which fares are likely to fare better with regard to survival (the color red denotes survived)?**

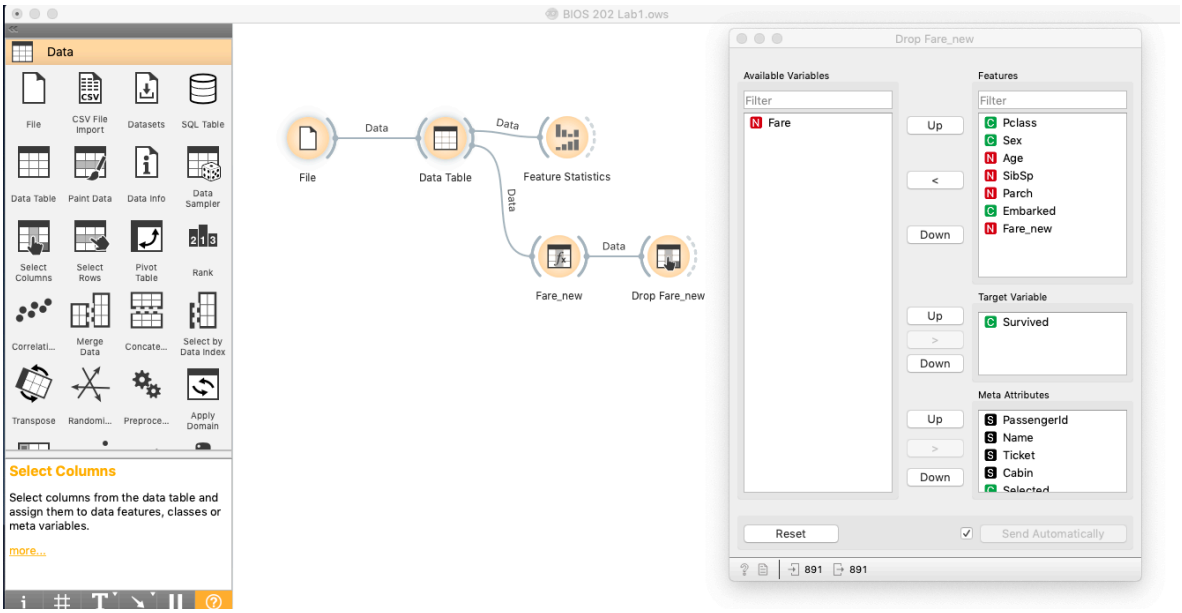


16. Consider the other columns of the "Feature Statistics" widget: Center tells you the central tendency of the variable (e.g. mean for numerical variables, mode for categorical variables), Dispersion tells you the spread of the data, Min. and Max. give the lower and upper values, and Missing tells you the proportion of data missing values for each variable. **Q3. Which variables have missing data? Look at the Min. value for *Fare*, does this make sense?** The "Feature Statistics" widget is a great tool to not only inspect the distribution and quality of your data but also begin to explore relationships between feature variables and the target variable.

17. The *Fare* variable has values of 0 which might be an error given that all tickets cost money. Therefore, we are going to find and replace all values of *Fare* = 0 and replace with a value of NaN which represents a missing value. To do so, connect a "Feature Constructor" widget to the "Data Table" widget (make sure all data is passed along the channel). We will create a new variable called *Fare_new* which will be the same as *Fare* with 0 values set to Nan. Name the widget "Fare_new" and open the widget. Set the tab New to Numeric. Replace the automatically generated variable named with the name *Fare_new*. In the Expression field, enter the following text: "Fare if Fare > 0 else nan" (see red box in screenshot below). This statement executes a logic that says, "Give our new variable *Fare_new* the same values as *Fare* except when *Fare* = 0 then set values to nan". Then click Send and close the widget. Connect a new "Data Table" widget and inspect the values. Are the 0 values in *Fare_new* missing? We will get a lot of experience using the "Feature Constructor" widget in the coming steps.

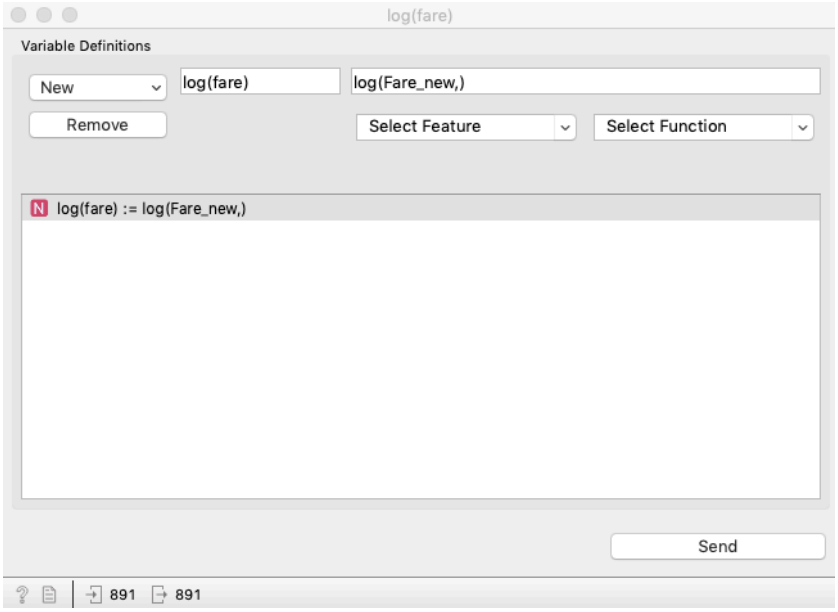


18. The next step is to drop the old *Fare* variable so it is not included in our analysis. Select the "Select Columns" widget, name it "Drop Fare" and attach to your "Feature Constructor" widget and open the widget. From the Features window, select *Fare* and hit the < arrow to move it into Available Variables. This will exclude the *Fare* variable from being passed through the channel. Close the widget.



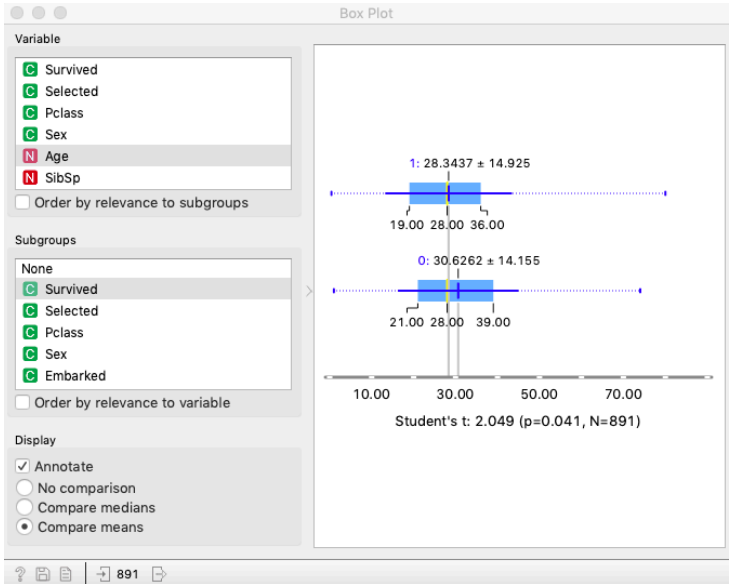
19. The *Fare* variable is skewed right but is likely to be an important variable for predicting survival. Sometimes with a highly skewed variable, predictive models will work better if they are transformed so as not to be so skewed. The logarithm of right skewed variables tends to be much less skewed, so we will create a log transformed version of the variable. To do so, add a "Feature Constructor" widget from the Data tab, rename to "log_fare" and connect it to the "Data Table" widget. Open the widget and from the New tab select Numeric, from the select

Function tab select log() and from the select Feature tab select *Fare_new*. This will create a new variable called *log_fare*. Click send. Note, in the future you can construct multiple features in a single widget rather than create separate widgets.

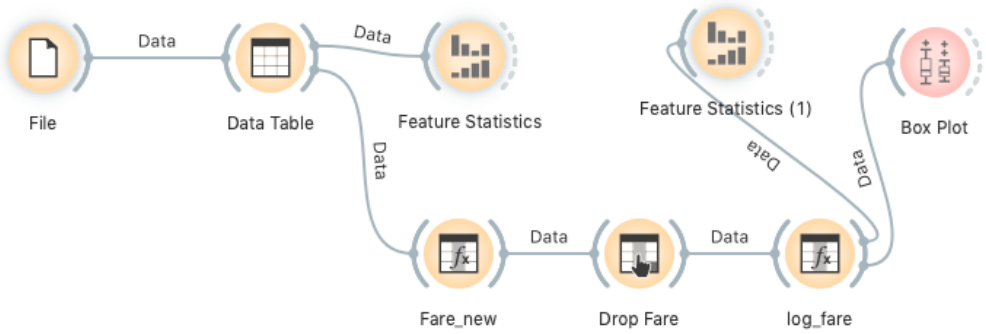


20. You can add a "Feature Statistics" widget and view the distribution of your new variable log_fare. **Q4. Is the log_fare variable less skewed? Does it have fewer outliers? Paste a screen shot of the histogram for the new variables.**

21. Next we will calculate some simple descriptive statistics. Let's create a plot that gives the average value of age and other descriptive statistics separately for those that did and did not survive. Under the Visualize tab connect a "Box Plot" widget to the "log_fare" widget. The "Box Plot" widget is extremely useful for stratifying feature variables across different levels or strata (I encourage you to play around). Open up the widget and select *Age* as Variable and *Survived* as Subgroups (see screenshot below). The box-and-whisker plot is stratified by *Survival* and tells us a lot of information. The vertical yellow bar and the shaded blue box represent the median (50^{th} percentile) and the interquartile range (25^{th} to 75h percentile). The blue vertical bar and horizontal blue line represent the mean and $\pm$ standard deviation. Finally, the dashed lines with blue border represent the range of the data. If the button Compare means is checked, we see the result of a statistical test comparing the means between subgroups. View the boxplot for *log_fare* and *Parch* by selecting from the Variable menu. **Q5. Does there seem to be differences between those that did and did not survive for these (hint, compare the mean values)? Do they make sense?** In the future, you can use a "Box Plot" widget along with a "Feature Statistics" widget to explore the data following import.



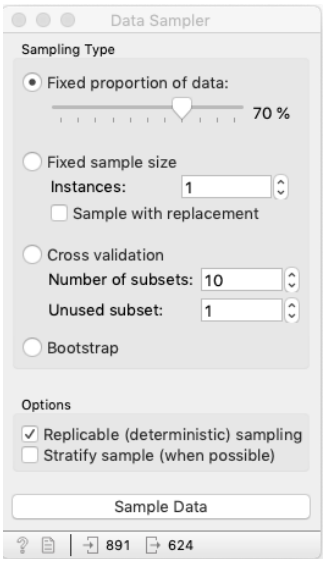
22. Housekeeping: When the workflows get complicated three things can help make them easier to read. First, you should name the widgets. Go back and rename your widgets as described above if you have not. Second, you can easily drag the widgets (or groups of widgets) around in the main window to arrange them more artfully to make the workflow easier to read. Your workflow by this point should look like this:



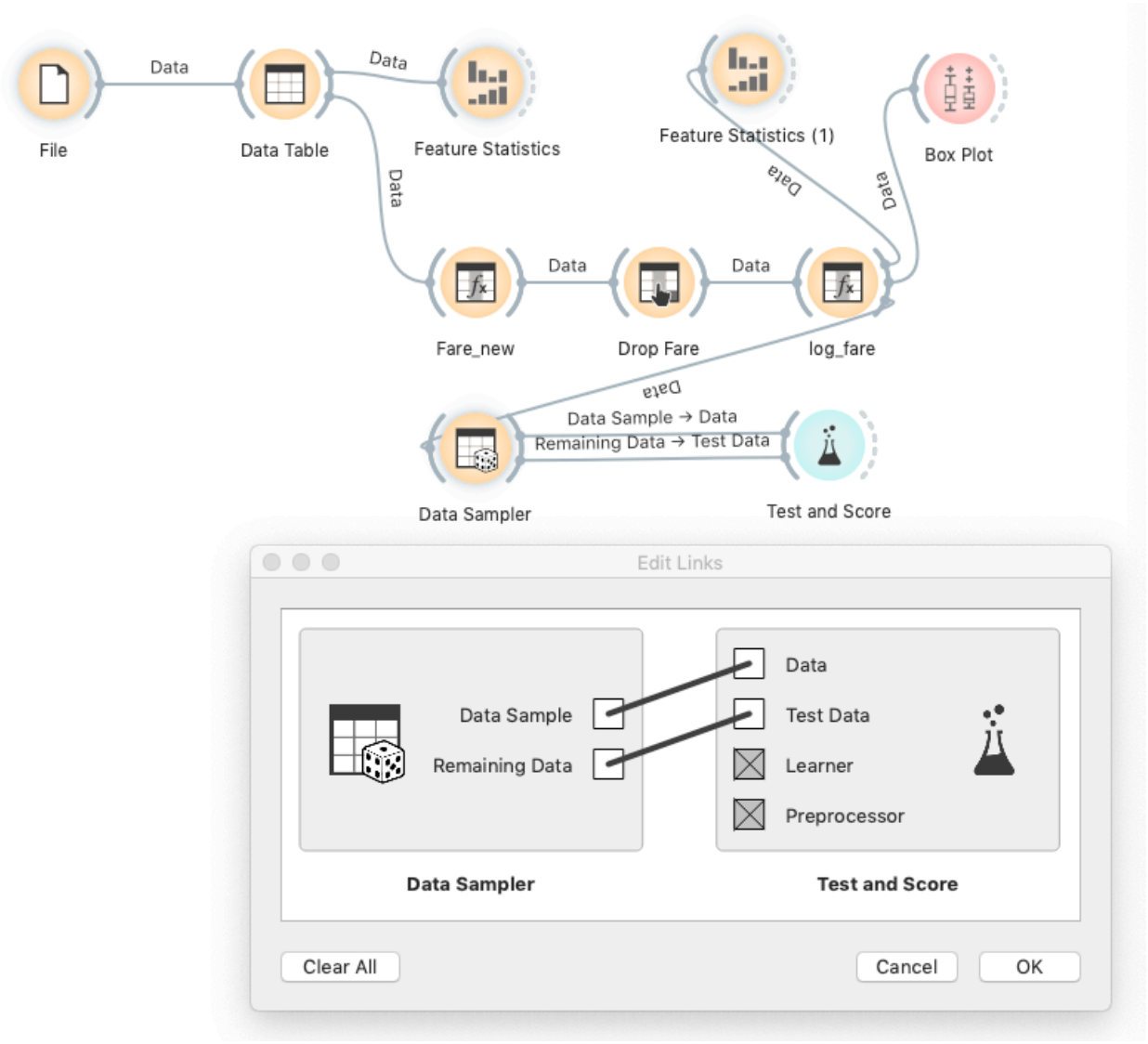
23. Now we are ready to build a model to predict who survived and who did not. A very good idea when building a predictive model (as will be discussed later in the course) is to reserve some of the data (called a “test” data set) to evaluate the performance of the model we developed (typically called “training” in the predictive modeling literature). If you try to use the same dataset to train and test a model, it usually appears over-optimistically good since it may adhere too closely to fluctuations in the data that just represent random variation that would not be present when using the model to predict additional cases.

24. The data can be randomly divided into training and testing subsets using a “Data Sampler” widget from the Data tab. Connect your “Data Sampler” widget to the “log_fare” widget. Open the “Data Sampler” widget and select Fixed proportion of

data to 70%. This divides the data randomly into two portions, one with 70% (training) of the data and one with 30% (test). There are other options we can discuss later. Keep Replicable (deterministic) sampling checked. This means we can reconstruct our partition at a later point perfectly. Then click Sample Data.



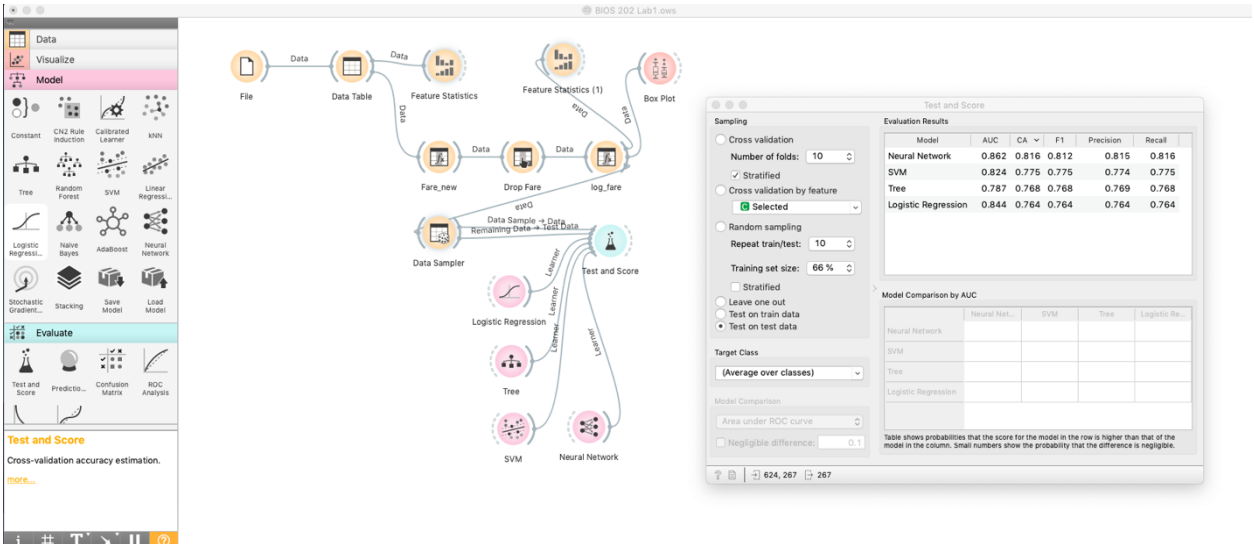
25. Now that we have partitioned the data into test and training portions, we need to tell Orange how to identify these data sets. From the Evaluate tab, select a “Test and Score” widget. Connect this to the “Data Sample” widget. Once the channel is established, click to edit. Connect the Data Sample from the “Data Sampler” widget to the Data of the “Test and Score” widget. Likewise, connect the Remaining Data from the “Data Sampler” to the Test Data of the “Test Data” widget. When we move to evaluate the performance of our model (test and score), this tells Orange exactly which data to train on and which to test on. Your channel configuration should look like this:



26. For our first model we will choose to run a logistic regression (more detail later in the course). It is a relatively standard statistical modeling method that takes multiple features (or predictors) to predict a categorical target variable (in this case survived or not). Add a “Logistic” widget from the Model tab to the “Test and Score” widget on the same side as the “Data Sampler” widget.

27. We are now going to evaluate our model performance. Open the “Test and Score” widget and under Sampling select Test on train data. Under the window Evaluation Results a number of numbers are listed next to Logistic Regression: AUC, CA, F1, Precision, and Recall. We will discuss these in depth later but for now focus on CA which is the accuracy (proportion of survival status correctly predicted). Note this number. Now under Sampling select Test on test data. Note this number. **Q6. What was the overall accuracy in the test and training datasets? Did it do better on the training or test dataset?**

28. Next try another modeling method to see if you can do better than the logistic regression, for example select one (or all) of the following widgets from the Modeling tab: “Tree”, “SVM”, or “Neural Network”. Add whatever Modeling wiget you select to the “Test and Score” widget. **Q7. What additional method did you try? Was it better, worse or about the same as the logistic regression when evaluated on the test data?** For the moment we have not explored the importance of particular features in predicting outcome, just the overall accuracy. In future labs, we will try to extract measures of which variables contribute most to the prediction of outcomes. My results are pasted below. They may not match yours exactly.



Q8: Final workflow (copy and paste a screenshot of your own workflow here).