MIS772

Predictive Analytics

Workshop: Model Deployment

Old process – New data, New process – Old data, New process – New data

DEAKIN BUSINESS SCHOOL

AACSB ACCREDITED

EFMD EQUIS ACCREDITED

Workshop Plan

Objectives:
The data set consists of 5000 reviews of air-travel and passenger recommendation of airlines based on their flight experience. Your task is to predict passenger recommendations.

Data Set:
Use files "airline-utf8-train.csv", "airline-utf8-valid.csv", and "airline-utf8-new.csv"

Acknowledgements:
The data has been "wrangled" by Quang Nguyen from Skytrax web site

Original Data:
https://github.com/quankiquanki/skytrax-reviews-dataset

Method:
Attend the seminar, follow the tutor's demo and instructions, take notes. Note that the class and online seminar will be recorded and their videos linked to the CloudDeakin topic for later access and study.

1 **Acquire data**
   (a) Load data and unzip
   (b) Read the "airlines" CSV files, and store

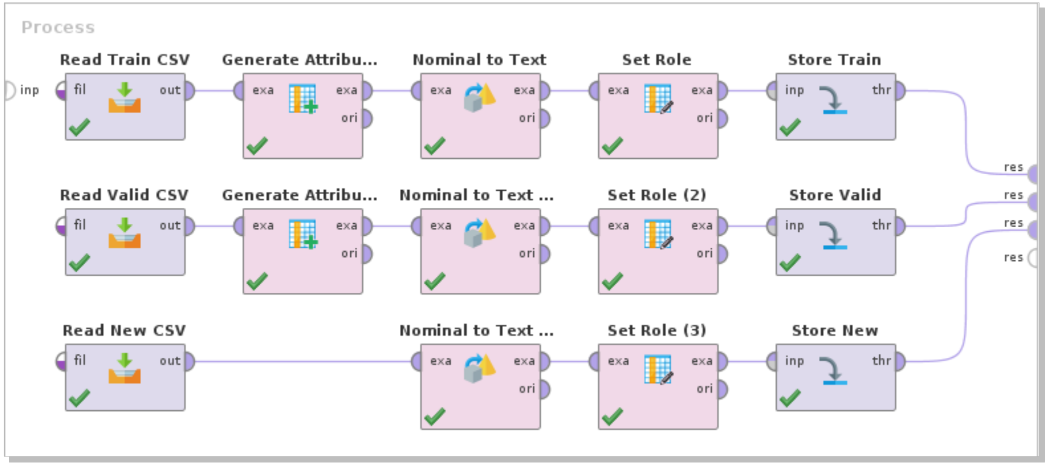
2 **Prepare data**
   (a) Select attributes and deal with missing values
   (b) Process text, reduce dimensionality and normalise
   (c) Bootstrap-validate logistic regression
   (d) Run, explore and save

3 **Create a simplistic model application**
   (a) Copy the training process for honest testing
   (b) Run, observe the result and save
   (c) What went wrong?

4 **Create a proper model application**
   (a) Modify the previous process
   (b) Transfer all pre-processing and predictive models
   (c) Transfer all word and attribute lists
   (d) Run, analyse and save

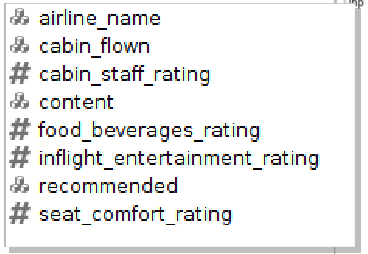
Data-Prep and First Model

Load Airlines data for training, honest testing and application. Redefine “recommended” to be true/false and set it as a label, convert “content” to text, also nominate “id” to be id, remember that “new” data has no label, store all data. Run and save.

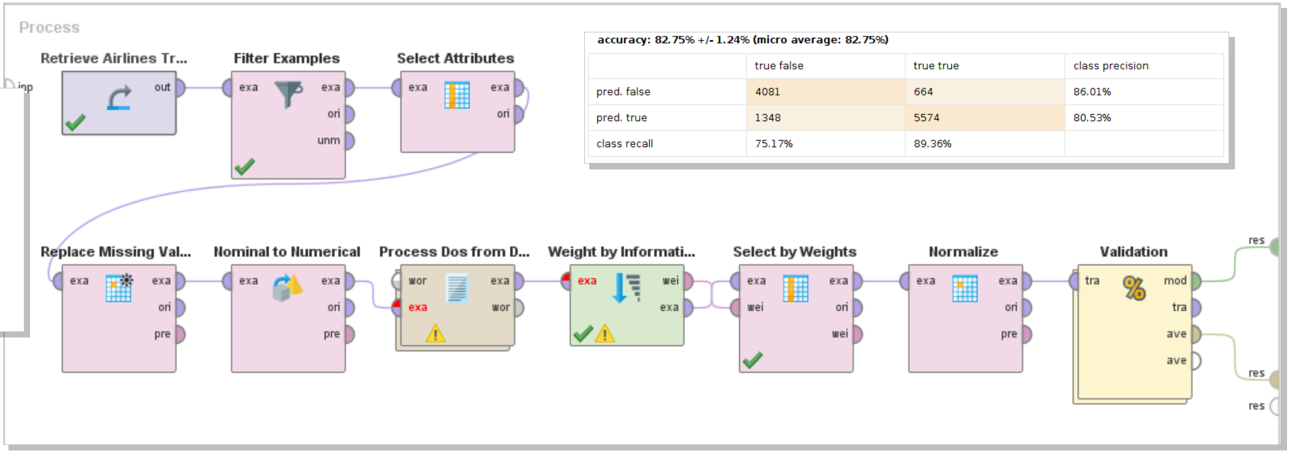


Create a CV Logistic Regression model with a mix of text and structured data. Filter out all missing labels. Select attributes for training (do not include overall rating). Deal with missing values, convert all nominal to numerical values (unique integers suggested). Process text. Select top 50 attributes. Normalise and bootstrap-validate the model. Note the model performance.

We will be bootstrap-validating the model, and we will be using the “validation” data set for honest testing



- airline_name
- cabin_flown
- cabin_staff_rating
- content
- food_beverages_rating
- inflight_entertainment_rating
- recommended
- seat_comfort_rating



accuracy: 82.75% +/- 1.24% (micro average: 82.75%)

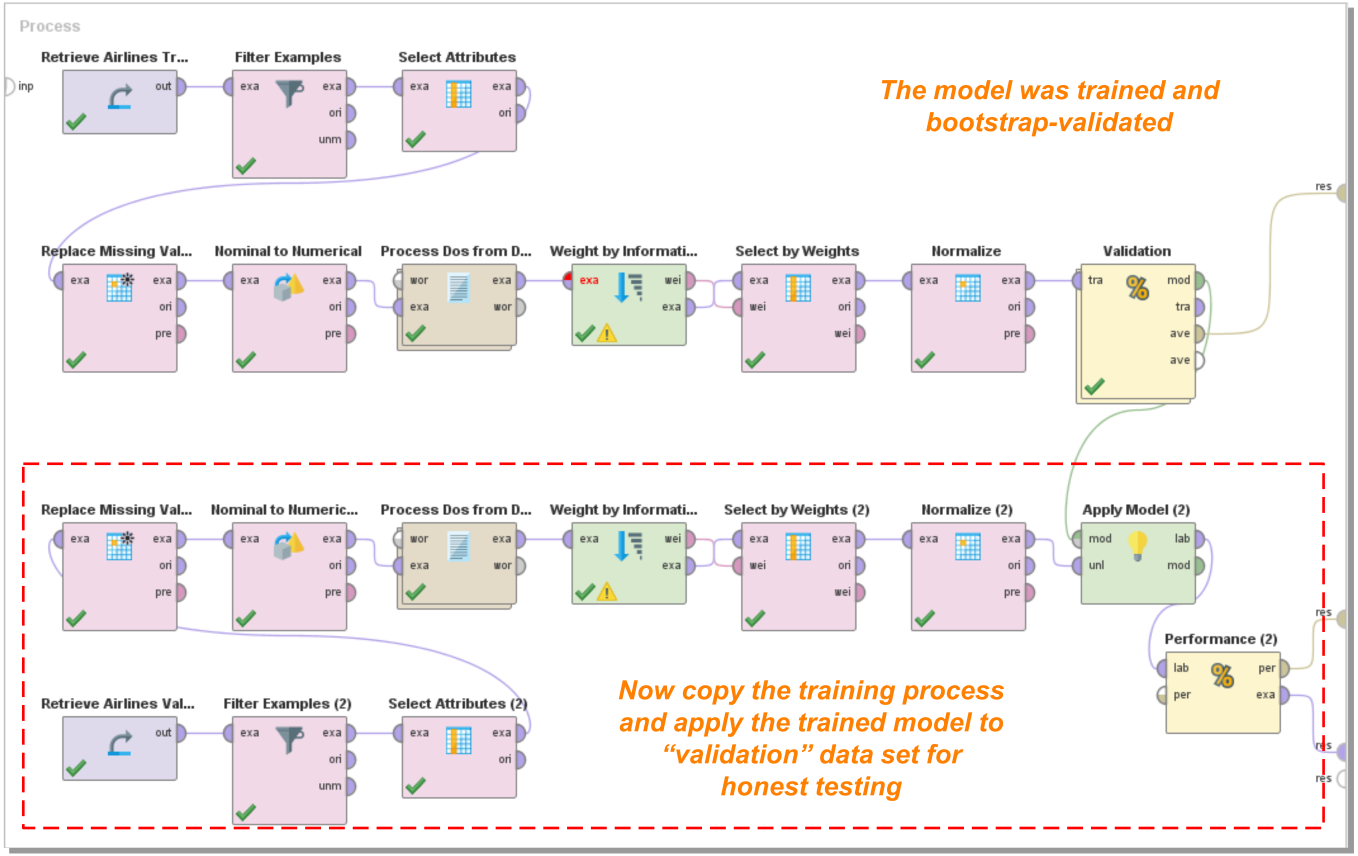
	true false	true true	class precision
pred. false	4081	664	86.01%
pred. true	1348	5574	80.53%
class recall	75.17%	89.36%	

DEAKIN BUSINESS SCHOOL

Simplistic “Application” of the model to new data

Copy the process and apply it to the “validation” data set for honest testing. What are you going to get? Why do you get this? What is wrong with this process? How should it be fixed?

Note that we have done all of this before, so this is just the consolidation of your previously learnt knowledge and skills



The model was trained and bootstrap-validated

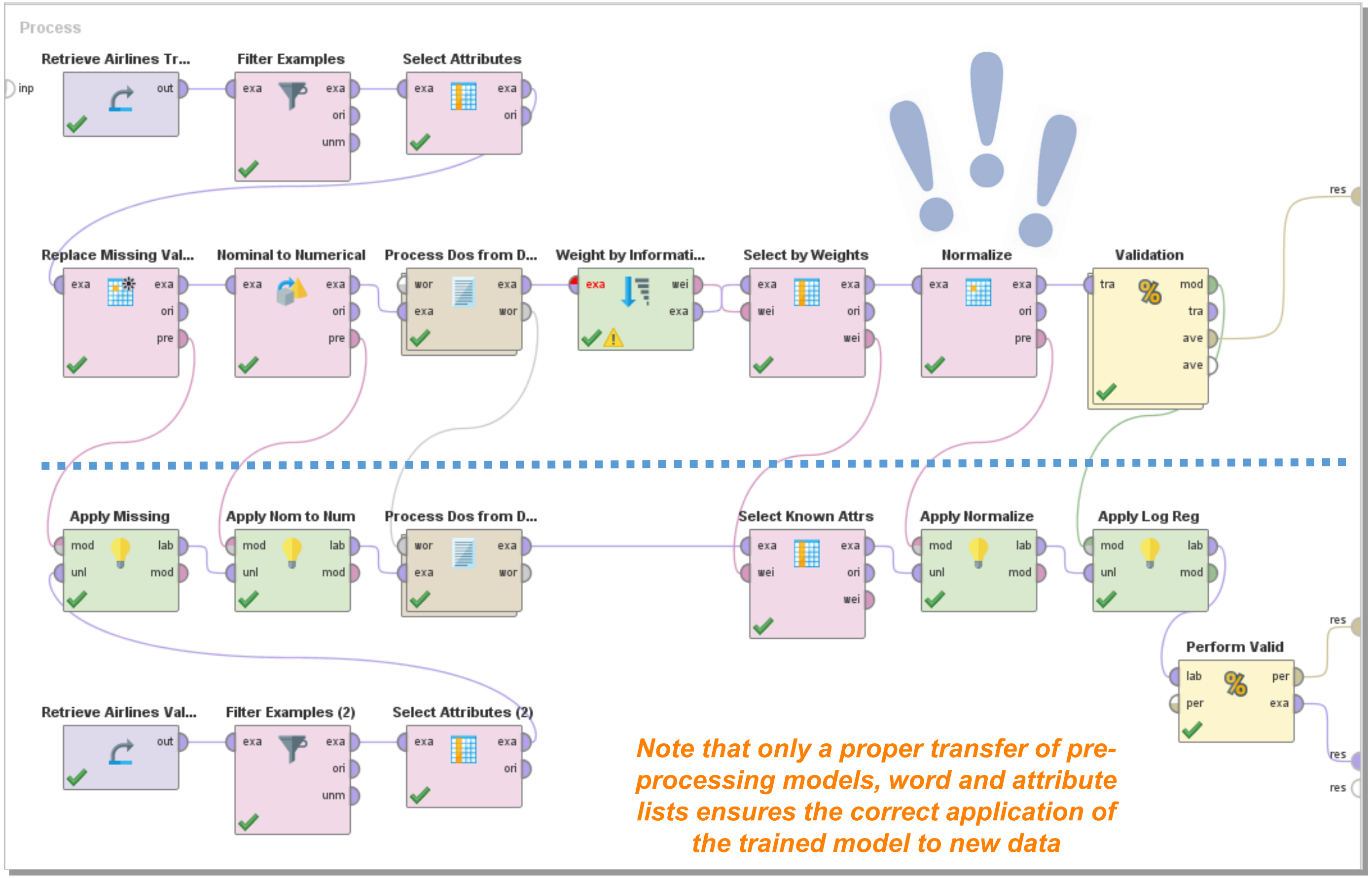
Now copy the training process and apply the trained model to “validation” data set for honest testing

DEAKIN BUSINESS SCHOOL

Improved application and deployment of the model

Correct the previous "bad" application of the model to ensure that all pre-processing and predictive models, word lists and attribute lists are transferred to the process honest-testing the model. What is the honest testing performance? (hopefully not as good as previously)

What aspects of the honest-testing processes model need to be transferred to for live deployment?



Note that only a proper transfer of pre-processing models, word and attribute lists ensures the correct application of the trained model to new data