Data, Algorithms and Meaning – Autumn 2020

Assignment 1 – Linear Regression and Classification Modelling

PART A – Linear Regression

This task will require you to develop and deploy linear regression modelling on a time series data set of financial transactions. A time series is simply a series of data points indexed in time order. In this data set, time is given in monthly intervals. The file provided for this section is **'transactions_monthly.csv'**. Below is a data dictionary for this file:

Field	Data Type	Description
date	Date	Date of the first day of each month
customer_id	String	Unique customer identifier
industry	Integer	Code for 10 industries, ranging from 1 to 10
location	Integer	Code for 10 locations, ranging from 1 to 10
monthly_amount	Numeric	Total transaction amount for customer in given month

You are working as a Data Scientist for a financial services company. The data set you've prepared describes the total transaction amounts for your customers each month. Transaction volumes can vary greatly for different product categories and locations, so these variables are included.

A sales manager wants to have an accurate prediction for *monthly_amount* next month.

Tasks:

Below are a series of tasks for you to undertake for Part A.

1. Undertake EDA on this dataset.
    a. Do you need to clean the data in any way? Justify what you decide to do (or not do).
    b. Describe two insights gained just from EDA that would be of interest to the sales manager.
2. Basic model fitting:
    a. Creating the model:
        i. Create an aggregated data set using the fields *date*, *industry* and *location*, with a mean of *monthly_amount.* ii. Create a line plot of the variable *monthly_amount* for *industry* = 1 and *location* = 1. Note the seasonality by month in this time series.
        iii. For *industry* = 1 and *location* = 1, train a linear regression model with *monthly_amount* as the target.

i. Note 1 :Remember that time is very important in this model, so be sure to include variable(s) for the time sequence. (Hint: on your plot you may see *local* trend like seasonality. Consider how you could craft a variable to capture this?. You may also see a global upwards or downwards slope, could you craft a variable to capture this? Therefore there are two simple variables you could create to capture time. Could you craft more complex ones, perhaps with polynomials to capture local or global trends? Experiment and see! ii. Note 2: Carefully think about how you split your test and train sets. (Hint: Random is not appropriate!)

iv. Create a prediction for *monthly_amount* in December 2016. Comment on how reasonable this prediction is. For example, if you were to plot it on the same plot as 2aii, would it sit somewhere reasonable?

b. Describe the model:
    i. How well does your model fit the data it is trained on in a statistical sense? Define & describe an appropriate quantitative measure. Justify your choice of measure.
    ii. How well does your model predicting *out-of-sample*? Define & describe an appropriate quantitative measure. Justify your choice of measure.

3. Advanced model fitting:
    a. Apply the modelling process you built for *industry* 1 and *location* 1 to all industries and locations programmatically.
    b. Calculate your evaluation measure for the training data and your testing data, for all models. Identify the two industries and two locations for which your method performs worst.
        i. Ensure your models all make a prediction for December 2016.
    c. What might be causing the models on these two industries and locations to be performing poorly (HINT: Some plots may help here…)? How might you fix this in future?
4. Reporting
    a. Using all the notes and answers you have above, wrap up all your work into a report for the sales manager that follows the CRISP-DM methodology. Whilst the sales manager is not a data scientist, they are intelligent and have some experience in data analytics. Therefore it will be an important task to ensure you include enough technical details to withstand QA and technical scrutiny whilst positioning for a business audience.
    b. Ensure you include your predictions on the test set and for December 2016 as an appendix. (Your predictions, the actual, the difference) which can be referenced in your report.
5. Submission.
    a. You must submit your report and professionally commented R-code.

PART B – Classification Modelling

This task will require you to develop and deploy a classification model on a product purchase data set.

You are now a Data Scientist working for an international consulting firm. An automotive manufacturer has approached you to help them target existing customers for a re-purchase campaign. The aim of this campaign is to send a communication to customers who are highly likely to purchase a new vehicle. All customers have already purchased at least one vehicle.

The automotive company has supplied a data set of customer demographics, previous car type bought, the age of the vehicle, and servicing details. Note that the servicing details are only for mechanics at official dealerships. You have looked at this data set and already transformed many of the variables to make the modelling easier. There was a lot of noise in the numeric variables so you decided to transform all the numeric variables into deciles (integers 1 to 10, each has a similar number of customers). The deciles can be treated as numeric or factors in R. The data dictionary for this data set is given below:

Field	Data Type	Description
ID	Unique ID	Unique ID of the customer
Target	Integer	Model target. 1 if the customer has purchased more than 1 vehicle, 0 if they have only purchased 1.
age_band	Categorical	Age banded into categories
gender	Categorical	Male, Female or Missing
car_model	Categorical	The model of vehicle, 18 models in total
car_segment	Categorical	The type of vehicle
age_of_vehicle_years	Integer	Age of their last vehicle, in deciles
sched_serv_warr	Integer	Number of scheduled services (e.g. regular check-ups) used under warranty, in deciles
non_sched_serv_warr	Integer	Number of non-scheduled services (e.g. something broke out of the service cycle) used under warranty, in deciles
sched_serv_paid	Integer	Amount paid for scheduled services, in deciles
non_sched_serv_paid	Integer	Amount paid for non scheduled services, in deciles
total_paid_services	Integer	Amount paid in total for services, in deciles
total_services	Integer	Total number of services, in deciles

mth_since_last_serv	Integer	The number of months since the last service, in deciles
annualised_mileage	Integer	Annualised vehicle mileage, in deciles
num_dealers_visited	Integer	Number of different dealers visited for servicing, in deciles
num_serv_dealer_purchased	Integer	Number of services had at the same dealer where the vehicle was purchased, in deciles

Tasks:

For this task you have been given two csv files, '**repurchase_training.csv'** and **'repurchase_validation.csv'**. The tasks below are to be undertaken on '**repurchase_training.csv'** since you have the target variable included.

1. Undertake EDA on this dataset. Justify all decisions you make relating to data processing.

2. Build a linear **classification** model to predict which customers are most likely to repurchase. You can use any technique we used in class with a classification target.
    a. Create the confusion matrix & calculate precision, recall, F1 & AUC.
    b. Given known best practice for classification tasks and the particular situation at hand, which metric will you use to decide your final model?

3. Build a tree based classification model to predict which customers are most likely to repurchase.

    a. How does your tree-based model perform relative to the linear model? (Hint: If you did not pick a good metric in 2.b you will not be able to compare the models!)
    b. Discuss the variable importance measures from both types of models (i.e what does 'variable importance' mean for these different model types).
    c. In this particular task, do they propose different levels of importance for the features? Why do you think this is? Provide intuitive interpretation of these importances, given the context at hand (i.e What do these importances mean in a 'business sense')
    d. For your tree-based model, construct partial dependency plots for the top 5 most important features. Do you notice anything interesting about any of these plots?

4. With the best model created above, use the variables in **'repurchase_validation.csv'** to output both probabilities and class predictions. You will see this file contains a validation

data set, but the target variable has been excluded. Your model will be marked against this data set. It is important to follow the following submission guidelines:

   a. name the file repurchase_validation_STUDENTNUMBER.csv
      i. So if your number is 1234 the file should be repurchase_validation_1234.csv
      ii. Include only three columns, named as follows, in this order. Note that target_class must be 1 for purchase and 0 for not. The probability here is for the positive (1) class. No need for two columns with two probabilities
         1. ID
         2. target_probability
         3. target_class

5. Reporting
   a. Using all the notes and answers you have above, wrap up all your work into a report for the manager of the automotive firm that follows the CRISP-DM methodology. Remember to consider your audience!