Data, Algorithms and Meaning – Spring 2017

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.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 industry
monthly_amount	Numeric	Total transaction amount for customer in given month

You are working as a Data Scientist for a financial services company. A data set has been prepared which describes the total transaction amounts for 10 of your customers each month. Transaction volumes can vary greatly for different industries and locations (coded from 1 to 10), so these variables are included.

A product manager has requested an accurate prediction for *monthly_amount* next month (December 2016), for each customer.

Tasks:

1. It is always best to start a modelling project by creating aggregated summaries and visualisations of your data.
    a. For each *customer_id*, explore the distribution of *monthly_amount* by calculated min, max, median, mean, and standard deviations. Also, examine the date range and number of observations for each customer.
    HINT: you can use the aggregate() function in base R to do this, or the summarise() function from the dplyr package).
    b. Create a line plot of the variable *monthly_amount* for some customers to get a feel for the data. Briefly describe the seasonality by month in the plots.
    HINT: You can take a sample of customer_id's and plot each time series separately, or write a for loop to loop over each customer, create a filtered data frame for that customer only, then plot it.
2. Choose a customer in the data set (any customer will do) and train a linear regression model with *monthly_amount* as the target. Remember that time is very important in this model, so be sure to include a variable for the time sequence (this can simply be a 1 for the first month, 2 for the second month, etc.). You can also use a regularised model if you wish.

   a. How well does your model fit the data it is trained on? Plot the line graph of the original data, and another line for your prediction *(HINT: you can use the lines() function to add another line)*. Define a quantitative measure to evaluate the goodness of fit.
   b. Describe how you can evaluate the accuracy of the model's predictions. Remember that this is predicting *out-of-sample* – for a time series the test data is usually a future period. *E.g., you can train the model on an initial period, then test on a later period that hasn't been used by the model.*
   c. Create a prediction for *monthly_amount* in December 2016 for the customer.
3. Apply the modelling process you built in section (2) to all customers.
   HINT: you can use a for loop as described in the FOR LOOP HINT below.
   a. Calculate your evaluation measure for the training data and your testing data, for all models. Identify the two customers for which your method performs worst, as well as the two for which it performs best.
   b. What might be causing the poor performance of the two worst models? How might you fix this in future?
4. Submit the data set with the model predictions for all customers for December 2016 in a separate csv file with the two fields *customer_id* and *prediction* only.

FOR LOOP HINT:
You can use this code template to create a for loop for building a model over all the customers. It is assumed you call your data set *data*. You will need to include additional code to partition your training and testing data.

```
evaluation_metrics = data.frame()
customers = unique(data$customer_id)
for (customer in customers) {
    temp = data[data$customer_id == customer, ]
    <DO ANY ADDITIONAL DATA PREPARATION HERE>
    model = <DEFINE YOUR MODEL HERE>
    temp$prediction = predict(model, temp)
    metric = <DEFINE YOUR EVALUATION METRIC HERE>
    evaluation_metrics = rbind(evaluation_metrics, metric)
}
```

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. Each row of data reflects a vehicle purchased. If a customer purchased two vehicles, the first purchase is flagged as a repurchase record (Target

= 1) and the second is not (Target = 0). For customers with more than 2 purchases, the last record has Target = 0 while all previous purchases have Target = 1. For records with Target = 1, all the data is calculated at the time of the second purchase. For records with Target = 0, the data is calculated at the current date.

You have looked at this data set already and transformed many of the variables to make the modelling easier. There was a lot of outliers and extreme values 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 lower decile numbers represent lower values of the variable. E.g. for *age_of_vehicle_years,* records with decile 1 had the lowest age while those with decile 10 had the highest. 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 record represents a repurchase, 0 if the record was the last or only purchase.
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:

1. Build a linear classification model to predict which customers are most likely to repurchase. You can use logistic regression or regularised regression with a classification target, as we did in class. Follow the CRISP-DM methodology given below. Train your model on data in the file **'repurchase_training.csv'**.

2. Build a tree based classification model to predict which customers are most likely to repurchase. Perform the same CRISP-DM steps as in task 1, but also discuss:
    a. How does your tree-based model perform relative to the linear model?
    b. Discuss the variable importance measures from both models. Do they differ between linear models and tree-based models?
    c. What do these models tell you about repurchasing customers, relative to single purchase customers? Discuss in light of predictor importance.

3. Choose a final model, either linear or tree based. Write a short report describing your approach, including choice of model, evaluation methodology and interpretation. Also describe any ethical or privacy issues involved with this task. For your final model, output both probabilities and class predictions to the file **'repurchase_validation.csv'**. This file contains a validation data set, but the Target variable has been excluded. Your model will be marked against this data set.

CRISP-DM Methodology: It is good practice to follow the CRISP-DM methodology when model building. In this instance, the business understanding step has already been applied to construct the data. From this point, data understanding is the first step, followed by data preparation, model building and evaluation. We repeat these steps until we have an optimal model. You should follow these steps in your modelling:

1. Understand your data by applying descriptive statistics and data visualisation.
2. Partition your data from the file **'repurchase_training.csv'** into a training set and a testing set.
3. Train your model on the training partition, and test on the testing partition.
4. Create probability and class predictions from your model. You may or may not need to decide on a probability threshold, depending on your model choice.
5. Create a confusion matrix. Calculate precision, recall and F1. From the output of this matrix, decide which model performs the best.

DAM Assignment 2 – Loan Default Model

Background

This task is to predict which loan customers will default on their repayments. All customer records provided relate to personal loans that have issued. There are a large number of predictors for the loan status target (either 'Fully Paid' or 'Charged Off'). The type of data provided and the modelling problem are commonly seen in the financial lending industry.

The data set contains a total of 42 columns and 39,786 rows. A description of all the columns and their values is provided in the file 'data_dictionary.csv'. The column 'loan_status' is the modelling target, and the other 41 columns are candidate predictors. Variations of the data set are publicly available, and a kaggle competition also uses this data:

https://www.kaggle.com/wendykan/lending-club-loan-data

You are free to explore previous work that has been completed on this data. However, caution is advised in using other people's solutions since the data provided in this assignment has been modified.

There are two key deliverables for this assignment, Part A and Part B.

Part A – Modelling

Work in groups to create a model to predict which customers are likely to have loan status of 'Charged Off'. The data and submission process are managed via a kaggle competition. There will also be a live leaderboard. The link to the competition is here:

https://kaggle.com/c/dam-spring2017-assignment2

There are two data sets, one training and one validation:

- training.csv
- validation.csv

You should train and test your model on the training data set, and run probability predictions on the validation data set. The performance of your model will be evaluated using the AUC measure (Area Under the ROC Curve) for binary classification models on the validation data set. One part of the validation data is public and will be made visible once you submit your predictions, but a second part is private and will be withheld until the assignment is finished.

Your team is allowed up to 100 submissions (you shouldn't need all of those), and up to 10 submissions per day.

Part B – Management Presentation

You are required to individually submit a management presentation on your approach. You should discuss the following, in line with CRISP-DM:

- The business problem
- The available data
- Your data preparation process
- Details of your model training, including the assumptions you made with a rationale
- Your evaluation methodology, including the model's performance as both a classifier (class predictions) and probability predictions
- Preliminary results (kaggle public evaluation measures)
- Any particular insights you discovered about the data
- Consideration of ethical issues

Your presentation should be short and concise, no more than 10 slides.