Xi'an Jiaotong-Liverpool University
西交利物浦大學

XJTLU Entrepreneur College (Taicang) Cover Sheet

Module code and Title	**DTS301TC Data Mining**
School Title	**School of AI and Advanced Computing**
Assignment Title	**Group Assignment**
Submission Deadline	**Sunday, October 16th 23:59 (Beijing Time), 2022**
Final Word Count	**N/A**
If you agree to let the university use your work anonymously for teaching and learning purposes, please type **"yes"** here.	

I certify that I have read and understood the University's Policy for dealing with Plagiarism, Collusion and the Fabrication of Data (available on Learning Mall Online). With reference to this policy I certify that:

- My work does not contain any instances of plagiarism and/or collusion.
  My work does not contain any fabricated data.

By uploading my assignment onto Learning Mall Online, I formally declare that all of the above information is true to the best of my knowledge and belief.

Scoring – For Tutor Use	
Student ID	

Stage of Marking	**Marker Code**	**Learning Outcomes Achieved (F/P/M/D) (please modify as appropriate)**			**Final Score**
		A	**B**	**C**	
1st Marker – red pen					
Moderation – green pen	**IM Initials**	The original mark has been accepted by the moderator (please circle as appropriate):			Y / N
		Data entry and score calculation have been checked by another tutor (please circle):			Y
2nd Marker if needed – green pen					

For Academic Office Use			**Possible Academic Infringement (please tick as appropriate)**	
Date Received	**Days late**	**Late Penalty**	☐ **Category A**	Total Academic Infringement Penalty (A,B, C, D, E, Please modify where necessary) _____
			☐ **Category B**	
			☐ **Category C**	
			☐ **Category D**	
			☐ **Category E**	

Xi'an Jiaotong-Liverpool University
西交利物浦大學

DTS301TC Data Mining

Group Assignment

Deadline: Sunday, October 16th 23:59 (Beijing Time), 2022.

Percentage in final mark: 40% (20% Group work + 20% Individual work)

Learning outcomes assessed: A, B, C

Late policy: 5% of the total marks available for the assessment shall be deducted from the assessment mark for each working day after the submission date, up to a maximum of five working days

Risks:

- Please read the coursework instructions and requirements carefully. Not following these instructions and requirements may result in loss of marks.
- The formal procedure for submitting coursework at XJTLU is strictly followed. Submission link on Learning Mall will be provided in due course. The submission timestamp on Learning Mall will be used to check late submission.
- Policies on Academic Integrity are strictly followed.

Overview

The purpose of this assignment is to get familiar with the basic concepts and techniques of data mining and gain experience in R and data mining applications. In this group project, you are expected to apply data mining techniques to predict car prices using the R programming language.

Dataset

The dataset used in this assignment contains information on around 18,000 used ford cars. In the *ford.csv* file, each record (row) contains information about a used ford car. The columns are explained as follows.

- model – the model of a car
- year – the registration year
- price – price in £
- transmission – type of gearbox
- mileage – distance used
- fuelType – engine fuel

- tax – road tax
- mpg – miles per gallon
- engineSize – size in litres

Requirements and Tasks

Given the datasets, you are expected to finish the following tasks using R programming language. You are allowed to use existing R libraries to solve the following tasks. Tasks 1 and 2 are group work; Tasks 3 and 4 are individual work. Please include all the source code and results for T1 and T2 in a group pdf file; include all the source code, results and evaluation report for T3 and T4 in an individual pdf file. Please also explain anything that is not obvious in the pdf files. Report clarity and brevity are valued over length.

T1. Exploratory Data Analysis – Group (25 marks)

T1-1: Load the CSV file; show the dimensionality, structure and summary of the dataset.

T1-2: Calculate and visualize the total amount of sales for each model.

T1-3: Calculate and visualize the total amount of sales per year.

T1-4: Calculate and visualize the average car price for each model.

T1-5: Calculate and visualize the average car price per year.

T1-6: Analyze data visualization results and summarize your findings in the pdf file.

T2. Data Pre-processing – Group (25 marks)

In task 2, you need to perform the following data pre-processing tasks on the given dataset. Each pre-processing task may be handled with different methods, e.g., fill or drop missing values. Please discuss with your team members and select a suitable method for those tasks.

T2-1: Check for missing values and handle them if they exist.

T2-2: Check for duplicates and remove them if they exist.

T2-3: Plot data distribution, check for outliers and remove them if they exist.

T2-4: Apply data normalization.

T2-5: Encode categorical values.

T2-6: Store the preprocessed dataset into a new CSV file.

T3. Modelling – Individual (30 marks)

T3-1: Each team member implements one different data mining model (e.g., kNN, linear regression, decision tree, random forest, SVM, etc.) to predict car price (attribute in the third column of the dataset) using the remaining attributes.

T3-2: Use k-fold cross validation with k = 5 folds to evaluate performance.

T3-3: Select features and/or tune model parameters to achieve the optimal performance. Show (or plot) model performance under different feature selection and/or parameter tuning settings.

T3-4: Report the best prediction results (i.e., RMSE, Rsquared and MAE) and the corresponding running time.

T4. Evaluation – Individual (20 marks)

T4-1: Use plots or charts of your model and one example to illustrate how your model is used for regression tasks.

T4-2: Discuss the performance of your model with your team members, i.e., RMSE, Rsquared, MAE and running time (Run the models under the same setting if necessary). Analyze the performance of your model.

T4-3: Discuss the advantages and disadvantages of the model you choose.

Group Submission

One group member must submit a zip file (named DTS301TC_ GroupID.zip) containing the following documents.

1. Cover letter with student IDs of all group members.
2. Source code files for Tasks 1 and 2.
3. A preprocessed dataset (in CSV format) generated in T2.
4. A pdf file containing all the source code and results for T1 and T2.

Individual Submission

Each student must submit a zip file (named DTS301TC_GroupID_IDNumber.zip) containing the following documents.

1. Cover letter with student ID.
2. Source code files for Tasks 3 and 4. Please name your source code file: IDnumber_YourName_ModelName.R (e.g.:1900000_ZhangSan_KNN.R).

3. A pdf file containing source code, results and evaluation report for T3 and T4. Please name your pdf file: IDnumber_YourName_ModelName.pdf (e.g.: 1900000_ZhangSan_KNN.pdf).