# ZC415 - Data Mining Assignment

1. Business Context

A big online retailer company “Amazing Zone” sells various products through their portal. One of the important feature of this business model is that the retailer takes responsibility of the shipment of the product. As the retailer takes guarantee that the purchased product will be delivered within promised timeline, customers are more inclined to purchase the products that has “Retailer fulfilled” tag associated with them. In order to keep the promise, the retailer has to invest heavily into the background shipment process that insures that the products are delivered to the customers before or on the guaranteed timeline.

1. Business Problem Understanding

The shipment process constitutes of several important components. The last mile delivery is carried out by the delivery boys, which are working on the company’s payroll. They are considered as critical because they are the one who faces the customer while delivering the product. They play an important role to increase the customer’s happiness by delivering the product before time, which improves upon the customer happiness index. These delivery personnel’s are many times over-loaded with the task of the deliveries, as company cannot afford to lose the customer because of bad impression at time of delivery. Due to the heavy workload, these delivery boys are not able to perform to their optimum limits on every business day, which results into their absenteeism from the work, which in turn increases the load on the other delivery personnel’s. This is not a good sign for the company as this influences the last mile delivery of the products.

You are the emerging analyst of data science department of company. You have been appointed to take a closer look at the absenteeism records of delivery personnel’s and

1. Identify the factors causing the absenteeism at work
2. Suggest a model that can help to determine the absenteeism hours for an employee
3. Suggest ways by which absenteeism can be reduced
4. Data Understanding

The data is gathered for three years absenteeism records of the delivery personnel’s. You need to make utilization of the features presented in the data set for your task. The data set and a document containing the information about the attributes are attached with the assignment problem statement. You can also visit the following link to access the data set and relevant information.

<https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work>

Make yourself familiar with these attributes as these might help you in determining the contributing factors with respect to the successful product purchase.

1. Data preparation and Exploratory Data Analysis

You are supposed to make utilizations of all the appropriate data pre-processing techniques on the given data set. If required, make appropriate assumptions and make it explicitly known while using them in the code or in the presentation. You are required to identify the key factors that influences the absenteeism at the work. Make appropriate selection of the attributes with sound justification for the same. The data set allows for several new combinations of attributes and attribute exclusions, or the modification of the attribute type (categorical, integer, or real) depending on the purpose of the research.

You are supposed to make use of Python programming language and its libraries to work on this analysis effort.

1. Model building and Evaluation

You are supposed to build a model that predicts the absenteeism hours for delivery personnel, provided the several features associated with the delivery personnel’s work are given as input.

Apply the appropriate evaluation techniques in order to determine the accuracy of the predictions made by the model. Think of employing the technique that helps in improving the accuracy of the models along with inclusion of limited number of factors in the model.

Try to obtain a model that can be easily understood and explained but it should not come at the cost of accuracy.

You are supposed to make use of Python’s scikit-learn library for this step. You are free to write your custom algorithm as well provided it help in trying the objective of the use case.

1. Expected Outcomes

The results should consists of

1. The python script file or Jupyter notebook containing all the code for the proposed solution. Write all code in single file only , with proper comments.
2. A presentation which describes the
   * business problem
   * your observations
   * suggestions to improve the situation at work

Convert the presentation in pdf format before submission.

Include these two documents within a zip file following the naming convention as :

“<Group\_No>\_<Group\_Leader\_Roll\_No>.zip”

## Evaluation Matrix

|  |  |  |
| --- | --- | --- |
| Sr No | Criteria |  |
| 1 | Data Understanding and Preparation along with EDA (40%) | * Data quality issues are identified and addressed * Derived attributes are identified and created * Any notable exceptions are reported in form of comments, wherever appropriate * Attempt in right direction to find out contributing factors * Right set of visuals are used for univariate and bivariate data analysis * Meaningful insights are derived and presented in effective manner |
| 2 | Model building and evaluation (40%) | * Right data mining task is identified * Train and test data derived and used properly * Appropriate data mining technique is used for the model building * Model parameters are fine tuned to improve the model accuracy * Appropriate technique is used to identify the factors contributing to the model accuracy * Model evaluation is done based on the appropriate measures and criteria’s |
| 3 | Effective Story telling through the presentation (10%) | * The presentation has proper structure, not too big, not too small. Elaborate the important points in more precise manner. * Has focus on the business problem to be solved * Talks about the factors contributing to the business issue along with right kind of proofs * Explaining the observations visually where visuals are showcasing the facts * The recommendations / suggestions are spelt out clearly. * Assumptions are specified at right places. |
| 4 | Code readability and organization (5%) | * Code is executing, no syntax errors * No customizations needs to be done in order to execute the code * Code is simple and augmented with proper comments wherever required * Built-in functions / libraries are used wherever possible * Repeated code is moved into functions and used appropriately when required * Long code snippets are broken down into small parts and made available as functions to increase the modularity of the code * Appropriate variable names are used to improve the readability of the code |
| 5 | Overall utilization of the concepts learnt in the course (5%) | * Appropriate steps are carried out in the data preparation stage which involves the concepts learnt in the class for the same * Various concepts from the data mining process are paid enough attention while developing the model |