# DSE ZG523 IDS Assignment 1 – PROBLEM STATEMENT 1

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 because of the traffic situations at the roads of the cities in which these deliveries are attempted. Due to these traffic situations, these delivery boys are not able to perform to their optimum limits on every business day. This is not a good sign for the company as this directly affects the last mile delivery of the products.

Currently the local distribution centers located in cities uses the static kind of scheduling technique to determine the delivery route for every delivery boy. That means, every morning the computer program prepares a delivery schedule for each delivery person which needs to be followed for completing the deliveries on the time. Many times, due to the traffic conditions on the roads, the delivery boys finds its challenging to complete the required number of deliveries which results into the delays in the whole process. To improve upon this behavior, the company is thinking of dynamic routing of the delivery personnel’s. That means, when they are on the roads, and if there are certain incident has happened on that route, then the boy will be informed about the same and also an updated route will be provided for him to follow.

You are the emerging analyst of data science department of company. In order to prepare for the above mentioned objective, you have been appointed to take a closer look at the records of behavior of the urban traffic of the city of Sao Paulo and

1. Identify the factors causing the traffic congestion
2. Suggest model that can help to improve upon the delivery schedule
3. Data Understanding

For this prototype, the database was created with records of behavior of the urban traffic of the city of Sao Paulo in Brazil from December 14, 2009 to December 18, 2009 (From Monday to Friday). Registered from 7:00 to 20:00 every 30 minutes.

A document (.arff file) containing the information about the attributes and the actual data values is attached with this assignment problem statement.



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. You are required to identify the key factors that influences the traffic congestions on the road. 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 two different models that predicts the slowness in the traffic for a given day for given time period, provided the several features 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. Don’t include data file in the zipped file.
2. A word document file containing answer to the following three sub questions based on the analysis that you have carried out earlier.
3. Without building any model, what are the contributing factors for traffic slowness?
4. Are you able to confirm the above findings using any two modelling techniques? Give appropriate explanation for the same.
5. Using Recursive Feature Elimination technique, what are the the contributing factors for traffic slowness?

## Evaluation Matrix

|  |  |  |
| --- | --- | --- |
| Sr No | Criteria |  |
| 1 | Data Understanding and Preparation along with EDA (55%)   * File Processing (20%) * Data Preprocessing (30%) * Identification of the factors causing the traffic slowness (5%) | * The given file is brought into appropriate format before doing any preprocessing * Data quality issues are identified and addressed * Derived attributes are identified and created * Any notable exceptions are reported in form of comments, wherever appropriate * Cleaned, ready to use dataset is provided * 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 (25%)   * Preparation and evaluation of two models (10 % each) * Identifying the suitable model for the given scenario (3%) * Identification of the factors causing the traffic slowness (2%) | * Right learning tasks are identified * Train and test data derived and used properly * Appropriate 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 | Using RFE technique to identify the factors contributing the slowness in the traffic (15%)   * Use of built in packages (3%) * Determining the factors using the custom code with adequate explanation (12%) | * Appropriate built in packages are used for RFE (Recursive Feature Elimination) * One by one features are eliminated from the model and its impact is explained * Significant features are identified using the model statistics without compromising much on the model accuracy between the iterations |
| 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 |

Additional References

1. Original Research Paper
2. [Scikit-learn Preprocessing Documentation](https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing)
3. [Scikit-learn Model selection and evaluation Documentation](https://scikit-learn.org/stable/model_selection.html#model-selection)
4. [Scikit-learn RFE Documentation](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html)
5. [RFE Case study](https://www.datacamp.com/community/tutorials/feature-selection-python)