### **Challenge Description**

### **StarAccessories Ltd**

StarAccesories Ltd is a dynamic start-up that manufactures accessories for mobile phones. Its customers are retailers worldwide. Lately, due to the rapid increase in the customer pool, the company needs to update and adjust its credit policy.

For this reason, data of several features have been gathered with the purpose to conduct data analysis to provide insights about on-time invoice clearance. You were hired to conduct this analysis and moreover, to build an ML model to analyze the predictability of these features on forecasting the on-time invoice clearance.

**Files**

train.csv - data used for training along with target variable

test.csv – data on which predictions are to be made

sample\_output.csv – sample format of submission

**Data Schema**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Type** | **Description** |
| UniqueId | Alphanumeric | *The unique id of an invoice.* |
| Region | String | *The geographical region of the customer.* |
| DigitizationOn | Date-String | *The date on which the customer proceeded to the digitalization of the office paperwork.* |
| InvoiceDate | Date-String | *The date on which the wholesaler charged the amount of the customer’s order. (Issued an invoice)* |
| Amount | Float | *The amount of the invoice* |
| DeadLinePaymentOn | Date-String | *The date on which the invoice should have been paid by the customer* |
| Avowed | Bool | *Whether the customer has accepted or not the invoice amount. True indicates the invoice is accepted while False indicates a disagreement.* |
| InvoiceType | Category-String | *The type of invoice can be of two types, ‘electronic’ or ‘paper’.* |
| Violation | Bool | *This is the* ***response variable****. True to indicate that the payment deadline was violated and False to indicate on-time clearance.* |

**Problem**

You are required to perform an analysis of the given data and learn how different features are related to and affect the outcome of violating the deadline. With the given data, build a machine learning model that can be used to predict whether a payment deadline will be violated.

For each record in the test set (test.csv), you must predict the value of *‘Violation’*. You should submit a CSV file with a header row and one row per test entry. The file (submission.csv) should have exactly 2 columns:

* **UniqueId**
* **Violation**

**Evaluation Metric:** The metric used for evaluating the performance of the classification model would be the **F1-score**. In [the statistical](https://en.wikipedia.org/wiki/Statistics) analysis of [binary classification](https://en.wikipedia.org/wiki/Binary_classification), the **F1 score** (also **F-score** or **F-measure**) is a measure of a test's accuracy. The F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

The metric is implemented under the **sklearn.metrics** module and is the [harmonic mean](https://en.wikipedia.org/wiki/Harmonic_mean#Harmonic_mean_of_two_numbers) of **precision** and **recall**:



For reading more about the F1-Score: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html>

**Deliverables**

* Well commented Jupyter notebook
* 'submission.csv'

Your notebook should contain your solution, visualizations, and thought process, including the top features that go into the model. If required, please generate new features. Make appropriate plots, annotate the notebook with markdowns, and explain the necessary inferences. A person should be able to read your Notebook and understand the steps are you taking and the reasoning behind them. **The solution would be graded on the basis of the usage of effective visualizations to convey the analysis and the modeling process.**