Term Project

611121203 張茗溱

* **The selected system/tool**

Apache Spark ML、Kafka

**Reason :**

Spark的機器學習庫（MLlib）提供了豐富的機器學習算法和工具，能夠處理大規模數據集。

Kafka 是一個高吞吐量的分佈式消息系統，特別適合用於大規模數據流的實時處理，將數據從生產者發送到pyspark應用程序，實現可擴展的實時數據處理。

* **The** **data** **set**

Health Insurance Cross Sell Prediction

在這個數據集中，目標是預測上一年度保單持有人對購買車輛保險的興趣。這包括性別、年齡、地區代碼、車輛屬性（年齡、損壞情況）等因素，目的是通過識別潛在客戶並相應地制定溝通策略，來增強公司的戰略規劃。

**Data Description**

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| id | Unique ID for the customer |
| Gender | Gender of the customer |
| Age | Age of the customer |
| Driving\_License | 0 : Customer does not have DL,  1 : Customer already has DL |
| Region\_Code | Unique code for the region of the customer |
| Previously\_Insured | 1 : Customer already has Vehicle Insurance  0 : Customer doesn't have Vehicle Insurance |
| Vehicle\_Age | Age of the Vehicle |
| Vehicle\_Damage | 1 : Customer got his/her vehicle damaged in the past.  0 : Customer didn't get his/her vehicle damaged in the past. |
| Annual\_Premium | The amount customer needs to pay as premium in the year |
| Policy\_Sales\_Channel | Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc. |
| Vintage | Number of Days, Customer has been associated with the company |
| Response | 1 : Customer is interested,  0 : Customer is not interested |

<https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction?select=test.csv>

* **Function** **Description**

1. 數據準備和模型訓練：

* 資料讀取：使用 PySpark 讀取訓練資料集 CSV 檔案。
* 特徵轉換：將類別特徵使用 StringIndexer 轉換為數值特徵。
  + 特徵組合：使用 VectorAssembler 將所有特徵組合成一個特徵向量。
  + 模型選擇：選擇並建立機器學習模型（例如LogisticRegression、RandomForest…等等）。
* Pipeline 構建：建立包含數據轉換和模型的 ML Pipeline。
* 模型訓練：使用training dataset進行模型訓練。
* 模型評估：使用validation dataset評估模型性能。

1. 模擬實時數據流：
   * + 使用 Kafka Producer 將testing data的每一筆資料作為消息發送至 Kafka 主題(test\_data)。

* 忽略標頭並確保資料格式與訓練數據相符。

1. 實時預測和結果回傳讀取：

* 使用 Kafka Consumer 讀取實時數據流。
* 將每一筆資料轉換成 PySpark DataFrame 格式。
* 將數據送入訓練好的模型進行預測。
* 將預測結果、機率等信息回傳至 Kafka，方便使用者可以訂閱讀取。

1. 模型更新與實時學習：

* 模擬新數據到來：使用validation data做模擬新資料料從Kafka讀取進
* 來加入原有訓練數據集，模擬新數據到來的情境。
* 更新 ML Pipeline：使用整體數據集重新構建並訓練 ML Pipeline。
* 持續學習：確保模型能夠在實時環境中持續學習，以適應新的模式和趨勢。
* **The** **reference**

[Learn How To Use Spark ML and Spark Streaming | by Navid Mashinchi | Towards Data Science](https://towardsdatascience.com/learn-how-to-use-spark-ml-and-spark-streaming-3a731485d052)