Capstone Projects for Big Data

Capstone Projects for Big Data

- The projects in general involve using Big Data technologies like PySpark components mainly PySpark SQL & PySpark MLLib as well as Hadoop components like Hive as required
- The datasets for each project typically will have 100,000+ rows in delimited file/files.
- The final deliverable for each case study consists of:
  - Detailed problem statement
  - Data dictionary describing all the columns of the dataset used
  - Solution approach listing all the tasks to be performed
  - The above will be presented as a MS-PowerPoint file
  - Solution code as a Jupyter notebook (.IPYNB) file with appropriate comments and documentation wherever required
  - Solution template with tips but not code to the students to help them to arrive at the solution and to implement it.

Hotel Reservations - Machine Learning model in PySpark

Problem Statement

- A large dataset with over a hundred thousands (a lakh) records of online hotel reservation s is provided.
- A significant number of hotel reservations are getting canceled for reasons like schedule conflicts or changes.
- Can we build a machine learning model to predict whether the customer is going to cancel the reservation or honor it?

Solution Approach

- Browse the dataset to understand the fields and any clean up to be done for example regarding the rows/fields with null values.
- Understand and perform the data pre-processing requirements such as string indexing, vectorizing, binarizing and/or bucketizing using the appopriate APIs.
- Build a model with the specified ML algorithm and get the evaluation metrics.

Hotel Reservations - Machine Learning model in PySpark

Data Availability

- Data is available as a flat file in delimited format
- Broad description of the data is given below

Data Description

- The file contains the different attributes of customers' reservation details :
  - Details provided by the customer such as Number and type of occupants, Room type, Start and end dates, Meal plans and so on
  - Other details added such as lead time, market segment, repeated guest or not, previous bookings, previous cancellations etc.

Hotel Reservations - Machine Learning model in PySpark

Sample records with header

hotel_reservations_sample.csv