DETAILED PROJECT REPORT

ANALYSE INTERNATIONAL DEBT STATISTICS

Submited By

Mr.Swapnil S.Pawshe.

Objective:

- Development of a predictive model for monitoring the World Bank's international debt data. The model will determine whether a countries economy improving or not.

Benefits:

- Observation about developing countries economy.
- Gives better insight of debt management system.
- Helps in easy flow for managing resources.

Data Sharing Agreement:

- Sample file name (ex fraud Detection 20062021_101010)
- Length of date stamp(8 digits)
- Length of time stamp (6 digits)
- Number of Columns
- Column names
- Column data type

Data Validation and Data Transformation:

- Power BI Desktop is an excellent tool for ETL (extract, transform and load) operations and can perform almost any transformation that its big brother, SSIS (SQL Server Integration Services) can.
- Most ETL operations can be performed with no code – this is a big plus for those new to Power BI or for rapid report development.
- I am using an open address data-set that has been intentionally modified to include various errors that we will identify and correct throughout the blog post. In the absence of an accompanying data dictionary I will use consistency as a measure of data quality.
- The process of data validation is iterative… **assess, clean, repeat**.

- With the help of power query we can access name, number and clean the same data. In number there can be postal code, street number, quantity of item etc.
- In name there can be person name, product name, region name, country name etc.
- We can also adjust the space ,date ,different sheets of same table, different files with data in power query etc.
- We can append different tables with different data types.

- Comma separated value files, often known as a .CSV, are simple text files with rows of data where each value is separated by a comma. These types of files can contain very large amounts of data within a relatively small file size, making them an ideal data source for Power BI.
- We can import file from Local files, one drive-business, one drive personal, SQL server, web, share point etc.
- We have imported file from local csv file.

- **Explore your data** - Once you get data from your file into Power BI, it's time to explore. Just right-click the new dataset and then click **Explore**.
- **Schedule refresh** - If your file is saved to a local drive, you can setup scheduled refresh so your dataset and reports in Power BI stay up-to-date. To learn more, see Data refresh in Power BI. If your file is saved to OneDrive, Power BI will automatically synchronize with it about every hour.

Model Training:

- Data Export from Local database: The accumulated data from Local database is exported in csv format for model training
- Data Pre-processing: We can also adjust the space ,date, different sheets of same table, different files with data in power query etc. We can append different tables with different data types.
- Check for null values in the columns. If present impute the null values.
- Encode the categorical values with numeric values.

Clustering –

Clustering is **an unsupervised machine learning algorithm that looks for patterns in data by dividing it into clusters**. These clusters are created such that the points are homogenous within the cluster and heterogenous across clusters.

- How we can use Power BI clustering?
- To cluster values, first select the Person column, go to the Add column tab in the ribbon, and then select the Cluster values option. In the Cluster values dialog box, confirm the column that you want to use to create the clusters from, and enter the new name of the column. For this case, name this new column Cluster.

- **K-Means clustering in Power BI uses the custom visual created by Microsoft**. This visual uses an R script in the back end to create the clusters. Download the custom visual here.
- **What are the examples of clustering?**

Here are 7 examples of clustering algorithms in action.

- Identifying Fake News. Fake news is not a new phenomenon, but it is one that is becoming prolific. ...
- Spam filter. ...
- Marketing and Sales. ...
- Classifying network traffic. ...
- Identifying fraudulent or criminal activity. ...
- Document analysis. ...
- Fantasy Football and Sports.

Prediction:

- The testing files are shared in the batches and we perform the same Validation operations data transformation and data insertion on them.
- The accumulated data from db is exported in csv format for prediction
- We perform data pre-processing techniques on it.
- KMeans model created during training is loaded and clusters for the pre-processed data is predicted
- Based on the cluster number respective model is loaded and is used to predict the data for that cluster.
- Once the prediction is done for all the clusters. The predictions are saved in csv format and shared.

Q & A:

Q1) What's the source of data?

The data for training is provided by the client in multiple batches and each batch contain multiple files

Q2) What was the type of data?

The data was the combination of numerical and Categorical values.

Q3) After the File validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder

Q4) How do you do logs in Power BI?

- The Power BI audit logs are available **directly through Microsoft Purview**. There's also a link from the Power BI admin portal: In Power BI, select Settings > Admin portal. Select Audit logs.

Q5) What techniques were you using for data pre-processing?

Removing unwanted attributes

► Visualizing relation of independent variables with each other and output variables Checking and changing Distribution of continuous values

► Removing outliers Cleaning data and imputing if null values are present.

► Converting categorical data into numeric values.

► Scaling the data

Q6) How training was done or what models were used?

- Before diving the data in training and validation set we performed clustering over fit to divide the data into clusters.
- As per cluster the training and validation data were divided.
- The scaling was performed over training and validation data
- Algorithms like SVM, XGBoost were used based on the recall final model was used for each cluster and we saved that model.

Q7) How Prediction was done?

The testing files are shared by the client. We Perform the same life cycle till the data is clustered .Then on the basis of cluster number model is loaded and perform prediction. In the end we get the accumulated data of predictions.

Q9) What are the different stages of deployment?

The deployment process lets you clone content from one stage in the pipeline to another, typically from development to test, and from test to production. During deployment, Power BI copies the content from the current stage, into the target one. The connections between the copied items are kept during the copy process.

Deployment pipelines enable creators to develop and test Power BI content in the Power BI service, before the content is consumed by users. The content types include reports, paginated reports, dashboards, datasets and dataflows