ANALYSIS, EXPLORATION AND PREDICTION OF US-CANADA BORDER CROSSINGS DATA

Group - Phoenix

Abdullah Ahmed
Master of Applied Computing
University of Windsor
ahmed5q@uwindsor.ca
110087175

Mohammed Zubair Ahmed
Master of Applied Computing
University of Windsor
ahmed8q@uwindsor.ca
110088299

Abdus Samee Mohammed
Master of Applied Computing
University of Windsor
mohamm9j@uwindsor.ca
110087708

Raj Manoj Dedhia
Master of Applied Computing
University of Windsor
dedhiar@uwindsor.ca
110088375

***Abstract*—**

The following analysis aims to discern the trends among the US-Canada border crossing data. This analysis can be used by different stakeholders at the port of entry in several ways. Infrastructure can be developed by focusing on the categories of travellers that are more expected at that port of entry. Local businesses can adapt their inventory and stocks if they are expecting many customers. Tourists can better plan their outings based on traffic. The tourism department can better adopt and prepare accordingly. Border Services can manage the workload if a lot of travellers are expected.

I. INTRODUCTION AND MOTIVATION

Cross- Border data is very essential to the economic growth of every country. Every sector, such as agriculture, customer services, retail and tourism rely on the influx of traffic at the port of entry. Numerous people cross the US-Canada border each month using various forms of road travel. Events like COVID-19, the US election, the Canadian election, and others have a significant impact on the number of travels each month. The analysis of this trend impacts majorly, the department of travel for both countries, local businesses, and transport in border cities.

The motivation behind this project is to assist the local businesses at the port of entries to better equip and prepare themselves based on the traffic in different months.

II. RELATED WORK

A study compares the number of Canadian cross shoppers versus Canadian domestic shoppers and how it impacts the economy of both US and Canada. The study extends the concept of out-shopping and claims that eighty percent of Canada's 34 million residents live close to the US border and make frequent trips depending on cheapness of various products. [1]

One study conducted by Sullivan and Kang in 1997 aimed to meet multiple objectives such as analysing the motivation of Canadian shoppers for cross-border shopping, discovering the characteristics of Canadian shoppers and examine the sources of shopping information [2]

Similarly, a study conducted by Lord, Petrevu and Parsa focus on the US/Canada cross border dining experience. Few other studies focus mostly on how the cross-border traffic can affect the currency rate of both the countries involved.

In contrast, this study aims to analyse the border crossing data and discover patterns among its which may be useful to the local retail businesses at the border based on the incoming traffic.

Using the data, we also build a model that has the capability to predict future trends of traffic in different months.

III. DATABASE AND DATASET

This section will cover the details of the database including the schema of the database.

A. MongoDB: MongoDB is a NoSQL document database which can be used to build large scale internet applications. It allows the user to build a scalable schema for storing the information. MongoDB stores the data in the database in a document in a format known as BSON (Binary JSON). The data is fetched from MongoDB in JSON format. MongoDB was chosen as the database project because of its high flexibility in the manipulation of data and its structure. MongoDB can also handle an extremely high amount of data due to its scale-out architecture. MongoDB has good integration with Python and PySpark; Hence MongoDB is chosen as the database.

B. DATASET

The dataset is obtained from Canada.ca official website which provides the inbound crossings at the US-Canada border at port level. The data reflects the number of vehicles travelling between US and Canada by trip characteristics, the length of stay in Canada, type of transportation and the port of entry.

IV. PROPOSED MODEL

The proposed application implements a prediction model for the number of vehicles crossing US-Canada Border.

Various steps are involved in creating the application:

1. Data- Pre-processing

Data- Pre- Processing in this application is done using Pandas and PyMongo library of Python.

Firstly, the required PyMongo and pandas libraries are imported, and the notebook is connected to the MongoDB cloud server.
A database is created using MongoClient() and then a collection called cleanedwithcoords is created.

The csv file of the dataset is read using pandas.read_csv method and converted to a dataframe for processing.

	REF_DATE	GEO	Trip characteristics	Length of stay	Mode of transportation	VALUE
0	1990-01	Yarmouth, Nova Scotia	Total United States vehicles entering	Same day	Automobiles	0.0
1	1990-02	Yarmouth, Nova Scotia	Total United States vehicles entering	Same day	Automobiles	0.0
2	1990-03	Yarmouth, Nova Scotia	Total United States vehicles entering	Same day	Automobiles	0.0
3	1990-04	Yarmouth, Nova Scotia	Total United States vehicles entering	Same day	Automobiles	0.0
4	1990-05	Yarmouth, Nova Scotia	Total United States vehicles entering	Same day	Automobiles	0.0
...	...	...	...	...	...	...
299947	2021-08	Whitehorse, Yukon	Total Canadian vehicles returning	Two or more nights	Automobiles	24.0
299948	2021-09	Whitehorse, Yukon	Total Canadian vehicles returning	Two or more nights	Automobiles	16.0
299949	2021-10	Whitehorse, Yukon	Total Canadian vehicles returning	Two or more nights	Automobiles	18.0
299950	2021-11	Whitehorse, Yukon	Total Canadian vehicles returning	Two or more nights	Automobiles	19.0
299951	2021-12	Whitehorse, Yukon	Total Canadian vehicles returning	Two or more nights	Automobiles	23.0

Fig 1. Dataset Dataframe

The columns unimportant for the model generation are dropped using df.drop.
The null values among the data are filled using the dataFrame.fillna() method.
Another column is created which consists of the REF_date data converted into datetime format. This datetime data is used by the Prophet model for training.
There were no values for the city of Windsor in the dataset from April 1997. To resolve this, the border crossing data for Windsor was imported from a similar dataset of border crossing data but generated by US government. The values were imported from that dataset to the main dataset.
The cleaned and pre-processed data is then inserted into the mongo DB database using PyMongo.

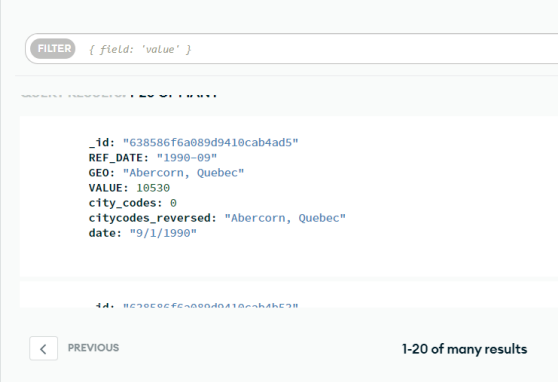

Fig 2. Collection in MongoDB

2. Building Model

The model is built using Prophet. Facebook's Prophet is an open-source forecasting tool which is optimized for forecasting tasks on data which has a strong seasonality characteristic, i.e., the data trends repeat after a particular period. At its core, the Facebook Prophet is an additive regressive model which fits non-linear growth with yearly seasonality with the help of Fourier series. It also takes into consideration trends such as holidays and is ideal to use with our dataset [3]

Training the model:
The date and value columns from the final dataset are taken as an input for the Prophet which trains the model. Next, we can predict the number of vehicles crossing the border using the model. The user can select a city and the number of months for prediction which will make a new dataframe containing the prediction dates for the selected city by using prophet's inbuilt method called make_futuredataframe. This dataframe is then used in the predict method of the prophet which will output the prediction

values. These prediction values are then used to plot a graph.

V. VISUALIZATION

A front-end user interface is created using streamlit framework which gives the user options to select the city and the number of months they want the prediction to be generated. The prediction values which were outputted from the trained model are then visualized in the user interface as graphs to comprehend easily.
The user is also presented with an option to view the historical data of any port of entry.

A map view of the North America is provided. The user can choose the year and view the number of vehicles crossing along the entire US-Canada border for that year.

VI. PROCESS FLOW

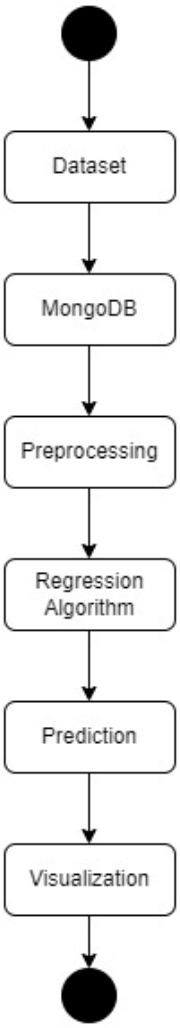

Fig 3. Process Flow

The process flow is as follows:

1. The dataset is downloaded from Canada.ca official website.
2. Dataset is stored onto MongoDB.
3. Data is fetched from MongoDB, cleaned, pre-processed, and stored in a collection.
4. The final pre-processed dataset is used as an input to the Prophet model which is a regression algorithm.
5. The model is trained, and the outputs are displayed.

VII. OUTPUTS/SCREENSHOTS

Fig 4. Historical Data

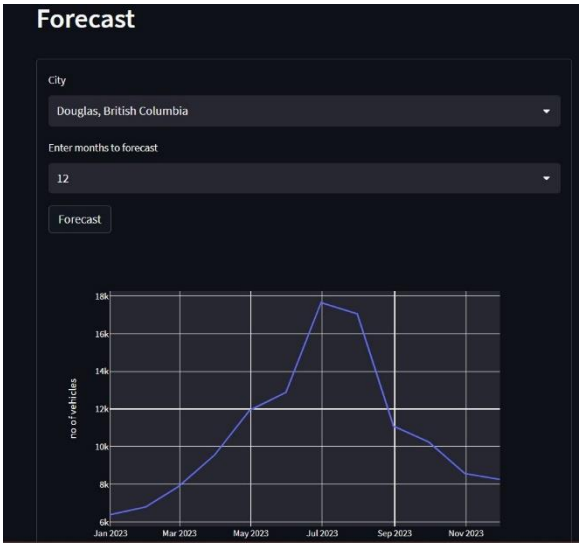

Fig 5. Forecasted Data for the next 12 months

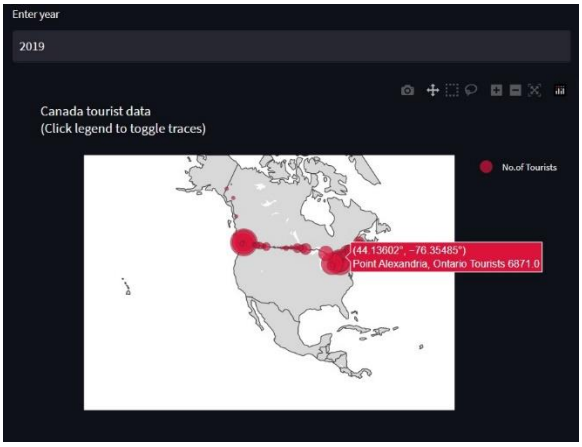

Fig 6. Map View

VIII. LIMITATIONS OR CHALLENGES

The model is not designed to handle anomalies like a sudden drop in the number of travellers. The model gave bad predictions when data from March 2020 onwards were included because the sudden reduction in numbers was not handled properly by the model and the model predicted negative number of travellers. This was solved by excluding the data from March 2020 onwards while training the model.

The dataset contained some columns that were unimportant for the training of the model and were hurting the accuracy of the model. The solution was to go by trial-and-error method and identify and drop the columns to improve the accuracy of the model.

IX. CONCLUSION

This study implemented a Time Series prediction model on US-Canada border

traveller's data using Facebook Prophet and predicted the expected number of travellers for any given port for the next 3/6/12 months. The suggested model enables the prediction of the influx of travellers based on previous trends with reliable accuracy. The predicted numbers correspond with the overall trend from a logical perspective and can be used referred assuredly.

X. FUTURE WORK

Future work would be to adjust the hyperparameters of the model and finetune it to improve the accuracy and overall prediction. The training data for the model can also be segmented based on the month before feeding it to the model which may lead to better accuracy of the model. Adjusting the model to handle anomalous data like sudden border closure due to COVID-19 can be a major improvement over the current implementation of the model.

XI. REFERENCES

[1] B. A. Zinser and G. J. Brunswick, "Cross-border shopping: A research proposal for a comparison of service encounters of Canadian cross-border shoppers versus Canadian domestic in-shoppers," *International Business & Economics Research Journal (IBER)*, vol. 13, no. 5, p. 1077, 2014.

[2] Z. Zhan, Z. Li, and M. Guo, "Research on data selection of cross-border retail e-commerce enterprises from the perspective of consumer search behavior—take AliExpress, a cross-border e-commerce platform, as an example," *Proceedings of the 5th International Conference on Financial Innovation and Economic Development (ICFIED 2020)*, 2020.

[3] Sean J. Taylor, Ben Letham, "Prophet: forecasting at scale", research.facebook.com. https://research.facebook.com/blog/2017/2/prophet-forecasting-at-scale/ (accessed November 5, 2022)

[4] "Streamlit docs," *Streamlit documentation*. [Online]. docs.streamlit.io (Accessed: 30-Nov-2022).