IMSE 586 - Big Data Analytics and Visualization

Project - Prediction with Supply Chain Data

Instructor - Fred Fang

Team 20
Sneha Anilkumar Kulkarni
Pranay Reddy Patlolla
Nikhil Gawate

Objectives

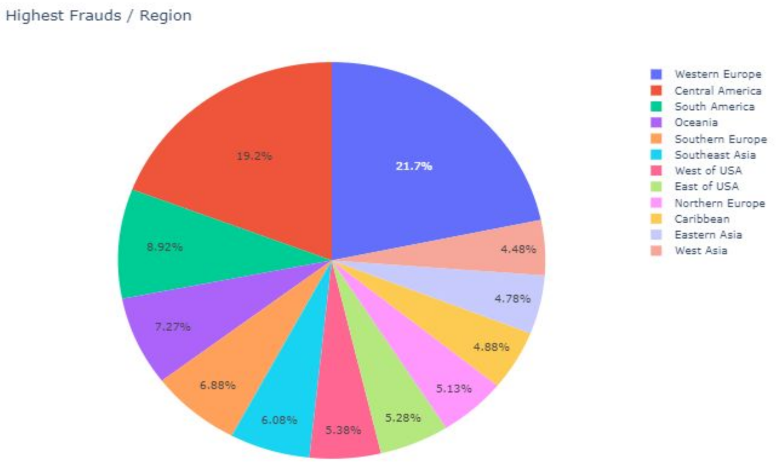
- Build a prediction model to predict late orders using the data which has the most impact on anticipated delivery date and fraud detection, Verify the model's accuracy in predicting the delivery date.
- Build a prediction model to predict complete orders using the data which has the most impact on financial performance.
- This dataset of supply chains, which is owned by the business DataaCo Global, contains information about the company's sold items, financial information, shipping information, and customer information, including sales, demographics, and transaction information. Contains approximately 180k transactions.

Approach

- Selecting a data set that would be helpful.
- Using previous supply chain data set consisting of the features which play a key role in prediction.
- Build the Pipeline and use part of the data to train the model.
- Cross validating the accuracy of models.
- Decision making.

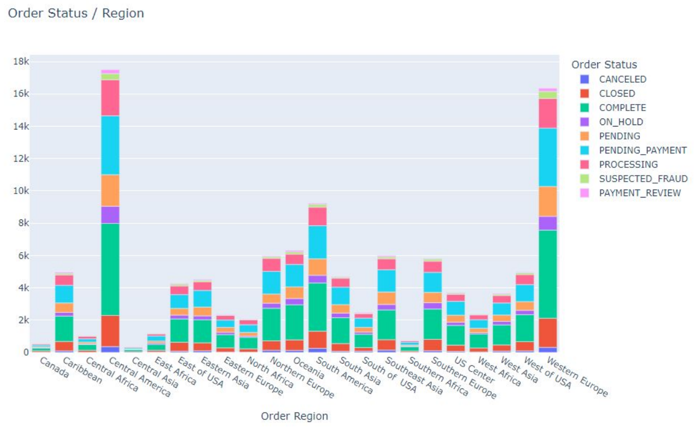
Data Analysis:

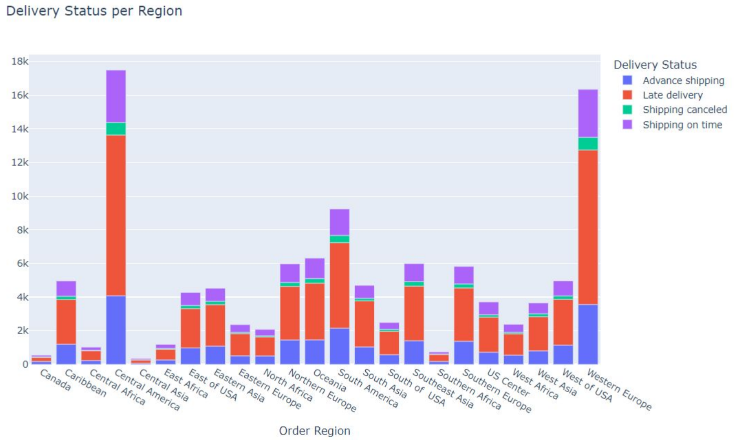
To have better insights about data we have done exploratory data analysis ands presented some visualizations.We have demonstrated the highest frauds occurring with respect to a particular region in the form of a pie chart.



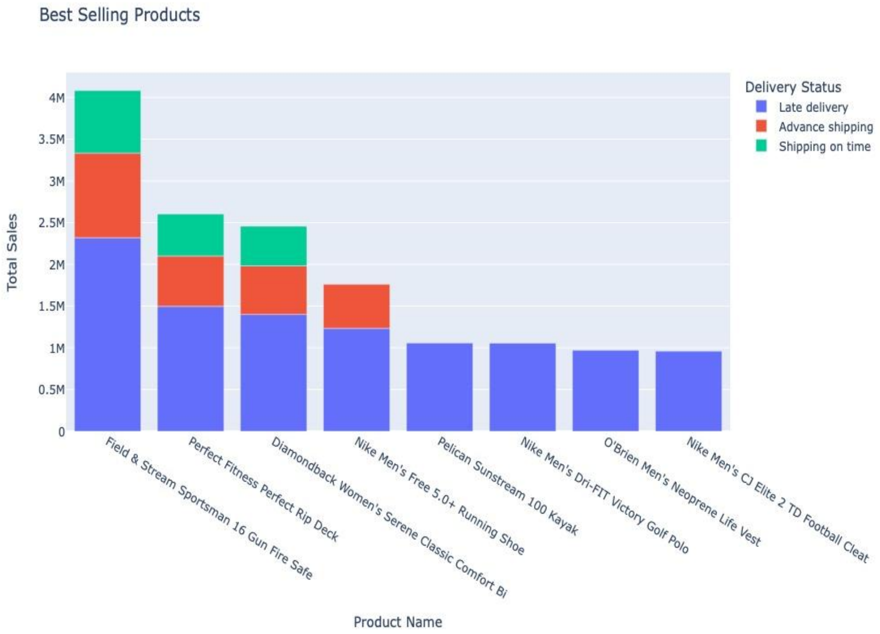
Delivery status & delivery status per region

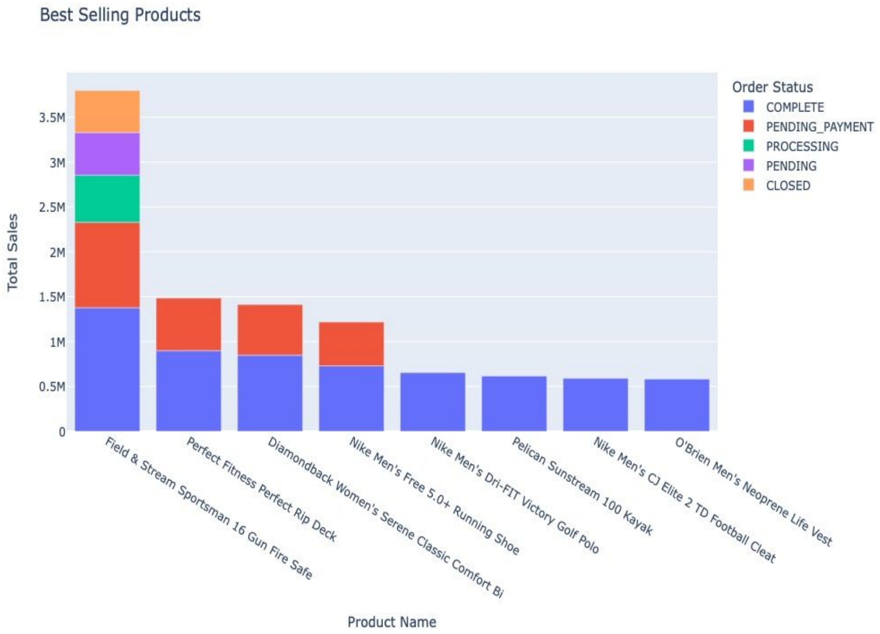
To have an overview about the order status with respect to country/region we plotted a bar graph.Similarly to view the delivery status per region:





Top 15 best selling products and delivery status



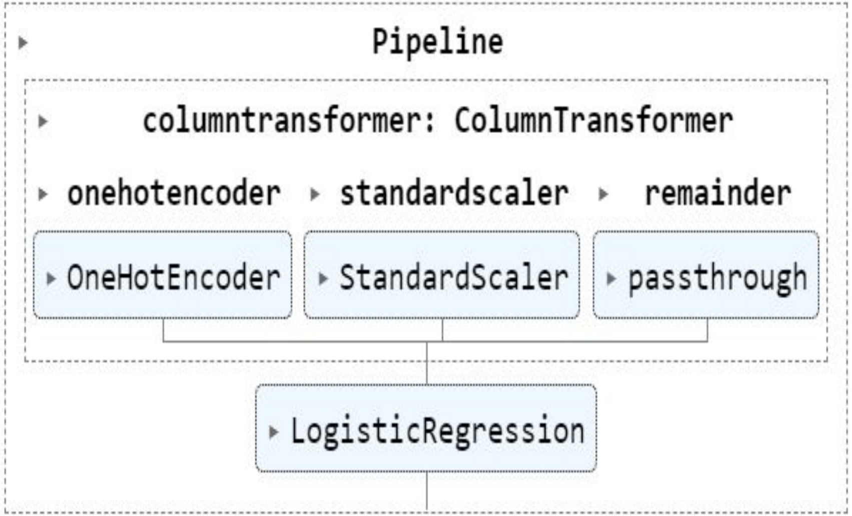


Data Modeling

For data modelling we have used the ML models :

- Logistic regression
- K-nearest neighbours
- Random Forest Classifier
- Naive Bayes Classifier

Logistic Regression

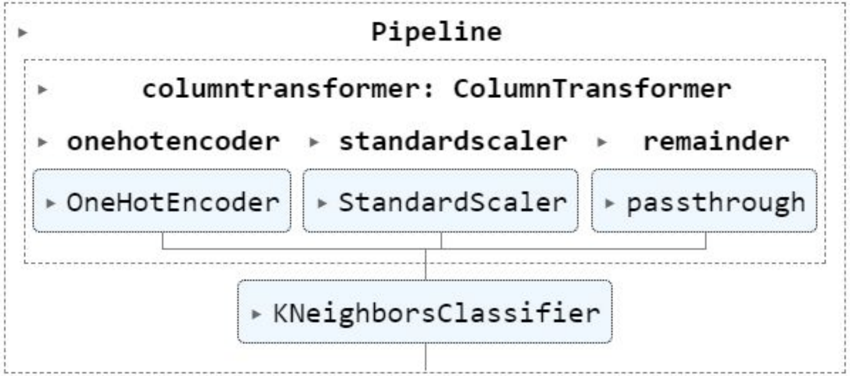


Accuracy: 0.9744

	precision	recall	f1-score	support
0	1.00	0.95	0.97	10059
1	0.96	1.00	0.98	12170
accuracy			0.97	22229
macro avg	0.98	0.97	0.97	22229
weighted avg	0.98	0.97	0.97	22229

	predict: Late delivery	predict: No Late delivery
actual: Late delivery	9515	544
actual: No Late delivery	26	12144

K-nearest neighbor

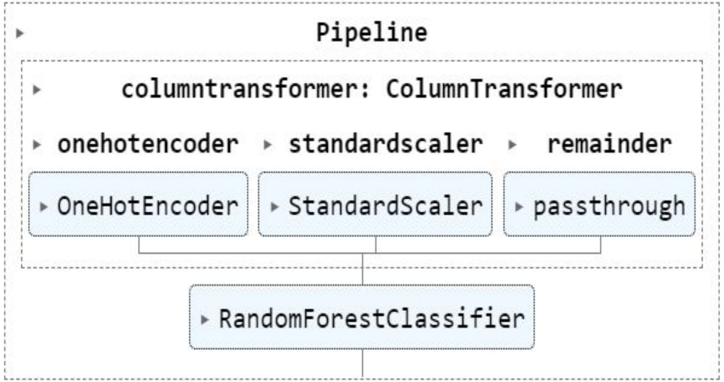


Accuracy: 0.9724

	precision	recall	f1-score	support
0	0.98	0.95	0.97	10059
1	0.96	0.99	0.98	12170
accuracy			0.97	22229
macro avg	0.97	0.97	0.97	22229
weighted avg	0.97	0.97	0.97	22229

	predict: Late delivery	predict: No Late delivery
actual: Late delivery	9597	462
actual: No Late delivery	151	12019

Random Forest Classifier

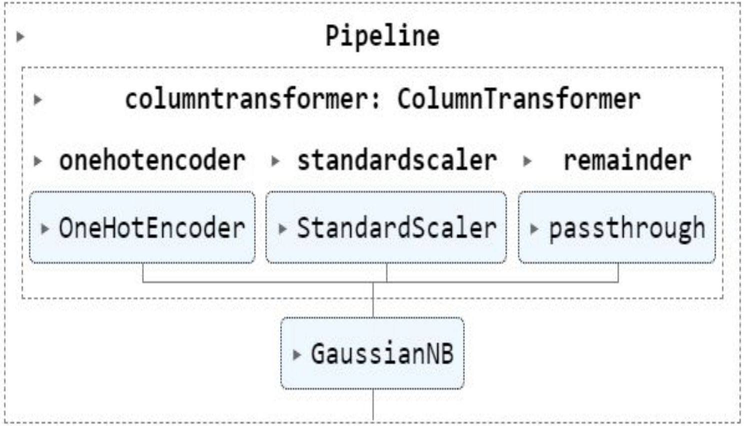


Accuracy: 0.9847

	precision	recall	f1-score	support
0	1.00	0.97	0.98	10059
1	0.97	1.00	0.99	12170
accuracy			0.98	22229
macro avg	0.99	0.98	0.98	22229
weighted avg	0.99	0.98	0.98	22229

	predict: Late delivery	predict: No Late delivery
actual: Late delivery	9730	329
actual: No Late delivery	11	12159

Naive Bayes Classifier



Accuracy: 0.7358

	precision	recall	f1-score	support
0	0.74	0.64	0.69	10059
1	0.73	0.81	0.77	12170
accuracy			0.74	22229
macro avg	0.74	0.73	0.73	22229
weighted avg	0.74	0.74	0.73	22229

	predict: Late delivery	predict: No Late delivery
actual: Late delivery	6479	3580
actual: No Late delivery	2294	9876

Limitations

- The original dataset when extracted from the source was found with the probability as shown below. As the percentage below represents, inside the dataset most of the transactions which were recorded are completed

```
In [5]: data['Order Status'].value_counts(normalize=True)

Out[5]: COMPLETE           0.329555
        PENDING_PAYMENT    0.220653
        PROCESSING         0.121328
        PENDING            0.112049
        CLOSED             0.108664
        ON_HOLD            0.054310
        SUSPECTED_FRAUD    0.022502
        CANCELED           0.020452
        PAYMENT_REVIEW     0.010486
        Name: Order Status, dtype: float64
```

- Unique values in Customer City and Order City

```
In [7]: data['Customer City'].nunique()

Out[7]: 563

In [10]: data['Order City'].nunique()

Out[10]: 3597
```

Results

The DataCo Company information was examined, and it was found that certain regions are found to have the highest number of fictitious transactions and orders with the most delayed deliveries.

We compared the accuracies of the models :

Random Forest Classifier	98.47%
KNearestNeighbours	97.24%
Logistic Regression	97.44%
Naive Byes	73.58%

Conclusion

- The Data Co company's orders with the risk of late delivery are delivered late every time.
- By the f1 score of 0.98, the Random forest classifier did a decent job of recognizing orders for later delivery and detecting fraudulent transactions.
- Hence implementing this model in the business for predicting late deliveries can be very helpful for the company.