Shipment Pricing Prediction

PARVEJ ALAM ANSARI

IBM Advanced Data Science Capstone

Use Case

- The market for supply chain analytics is expected to develop at a CAGR of 17.3 percent from 2019 to 2024, more than doubling in size.
- This data demonstrates how supply chain organizations are understanding the advantages of being able to predict what will happen in the future with a decent degree of certainty.
- Supply chain leaders may use this data to address supply chain difficulties, cut costs, and enhance service levels all at the same time.

Dataset

- The dataset was obtained from USA Government Website

(https://data.usaid.gov/HIV-AIDS/Supply-Chain-Shipment-Pricing-Data/a3rc-nmf6)

- Dataset consists of 10324 samples and 33 features of which 26 are categorical and 7 are numerical.

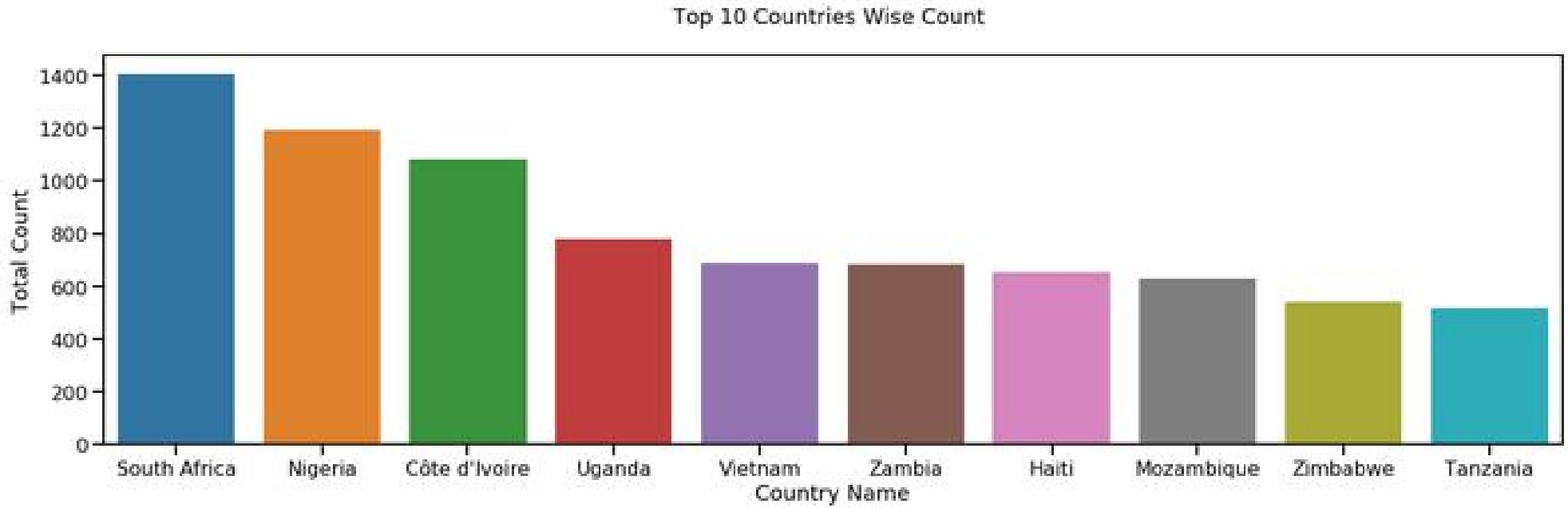
```
In [6]: 1 df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10324 entries, 0 to 10323
Data columns (total 33 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   ID                             10324 non-null  int64
 1   Project Code                   10324 non-null  object
 2   PQ #                           10324 non-null  object
 3   PO / SO #                      10324 non-null  object
 4   ASN/DN #                       10324 non-null  object
 5   Country                        10324 non-null  object
 6   Managed By                     10324 non-null  object
 7   Fulfill Via                    10324 non-null  object
 8   Vendor INCO Term               10324 non-null  object
 9   Shipment Mode                  9964 non-null   object
 10  PQ First Sent to Client Date   10324 non-null  object
 11  PO Sent to Vendor Date         10324 non-null  object
 12  Scheduled Delivery Date        10324 non-null  object
 13  Delivered to Client Date       10324 non-null  object
 14  Delivery Recorded Date         10324 non-null  object
 15  Product Group                  10324 non-null  object
 16  Sub Classification             10324 non-null  object
 17  Vendor                         10324 non-null  object
 18  Item Description               10324 non-null  object
 19  Molecule/Test Type             10324 non-null  object
 20  Brand                          10324 non-null  object
 21  Dosage                         8588 non-null   object
 22  Dosage Form                    10324 non-null  object
 23  Unit of Measure (Per Pack)     10324 non-null  int64
 24  Line Item Quantity             10324 non-null  int64
 25  Line Item Value                10324 non-null  float64
 26  Pack Price                     10324 non-null  float64
 27  Unit Price                     10324 non-null  float64
 28  Manufacturing Site             10324 non-null  object
 29  First Line Designation         10324 non-null  object
 30  Weight (Kilograms)             10324 non-null  object
 31  Freight Cost (USD)             10324 non-null  object
 32  Line Item Insurance (USD)      10037 non-null  float64
```

Data Quality Assessment

- In Dataset, outof 33 features, 3 features contain missing values.
- Some duration values were negative suggesting outliers or missing values
- The numerical features are highly skewed and categorical features needs to be properly handled.
- There were no Duplicates in the dataset.

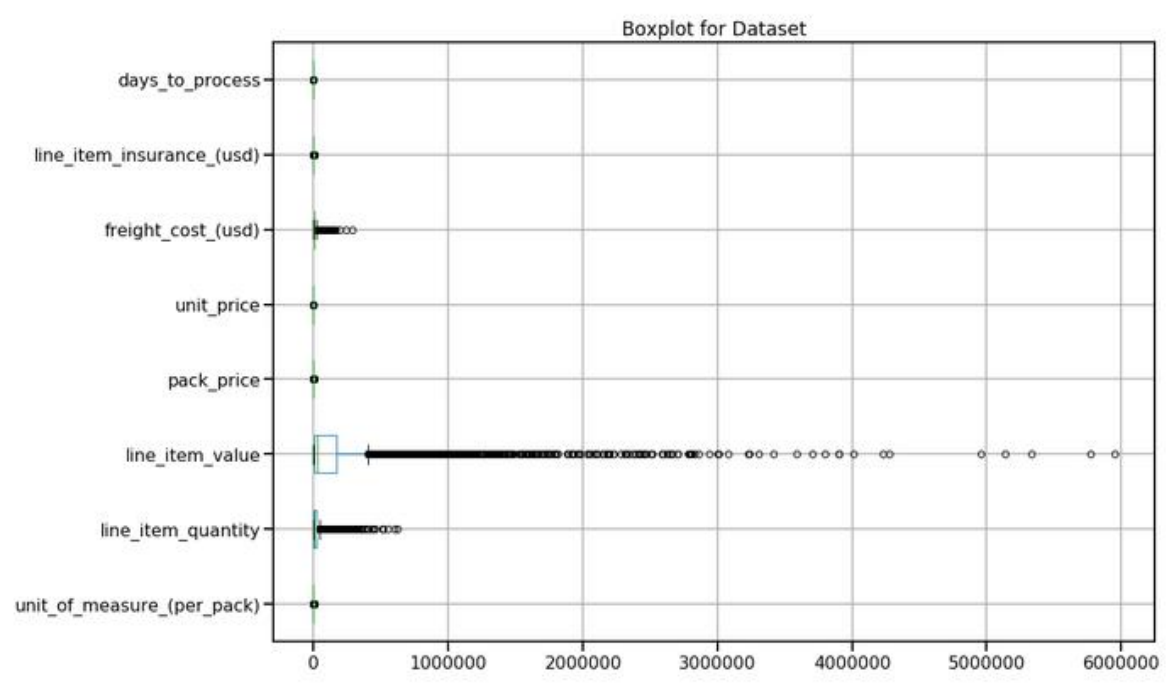
Exploration and Visualization

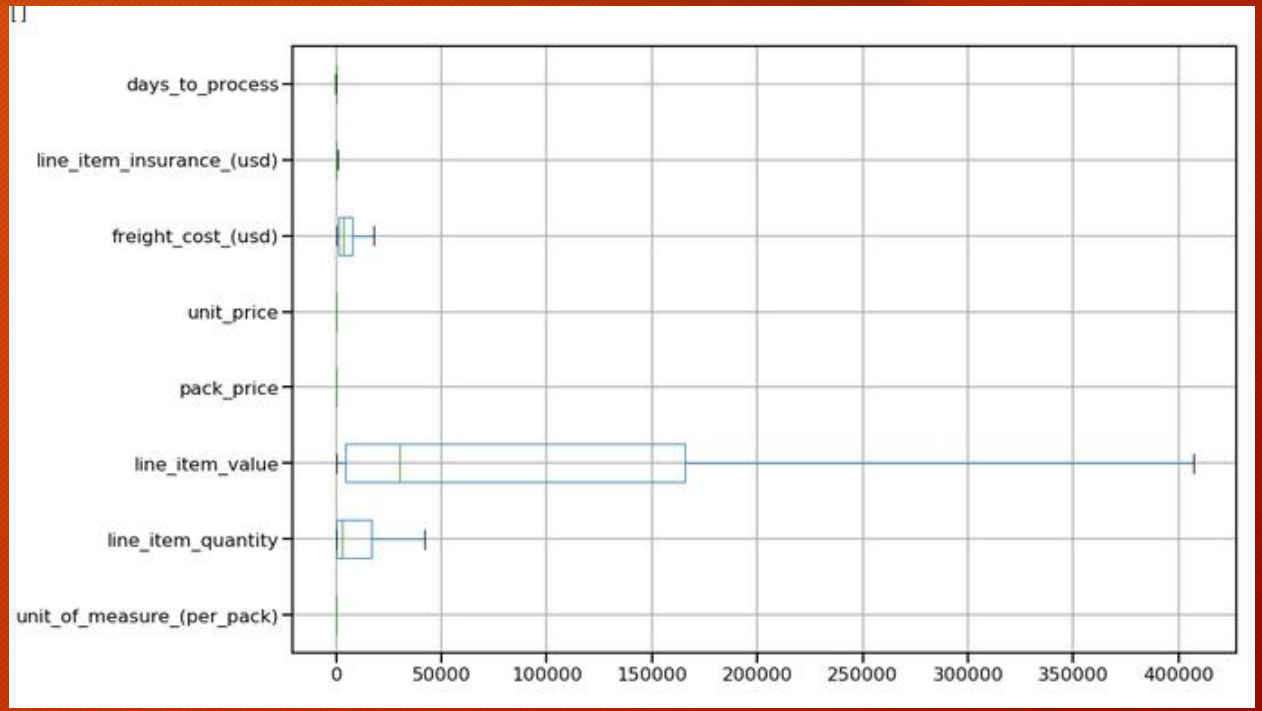


- South Africa is the most preferred Destination country for the shipments followed by Nigeria , Côte d'Ivoire while Guinea and Belize being the least.

South Africa and Nigeria are most preferred Destination country for shipments.

Exploration and Visualization





Enhanced Box Plots for outlier Detection

Feature Engineering

- Missing values were properly handled as pre the need.
- Each numerical values was clipped to remove outliers.
- Categorical variables were converted to Numerical one for processing.
- Feature Scaling was performed using MinMax scaler.
- Numerical features and Categorical features were preprocessed for better performance of the ML Algorithms.

Model Selection

- Two models were uses:-
  1. Gradient boosted tree implemented with LightGBM
  2. Deep Neural Network implemented with Keras
- Hyper-parameters For each models were optimized using Bayesian optimization
- The models were evaluated on all 3 three feature sets.

Results

Out[224]:

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	30305.201637	30193.461306	0.952646	0.950245
Decision Tree Regressor	13334.788570	22436.725747	0.990832	0.972526
Random Forest Regressor	16430.245975	18590.446875	0.986081	0.981138
ANN Regressor	20613.034324	19750.531817	0.978092	0.978711

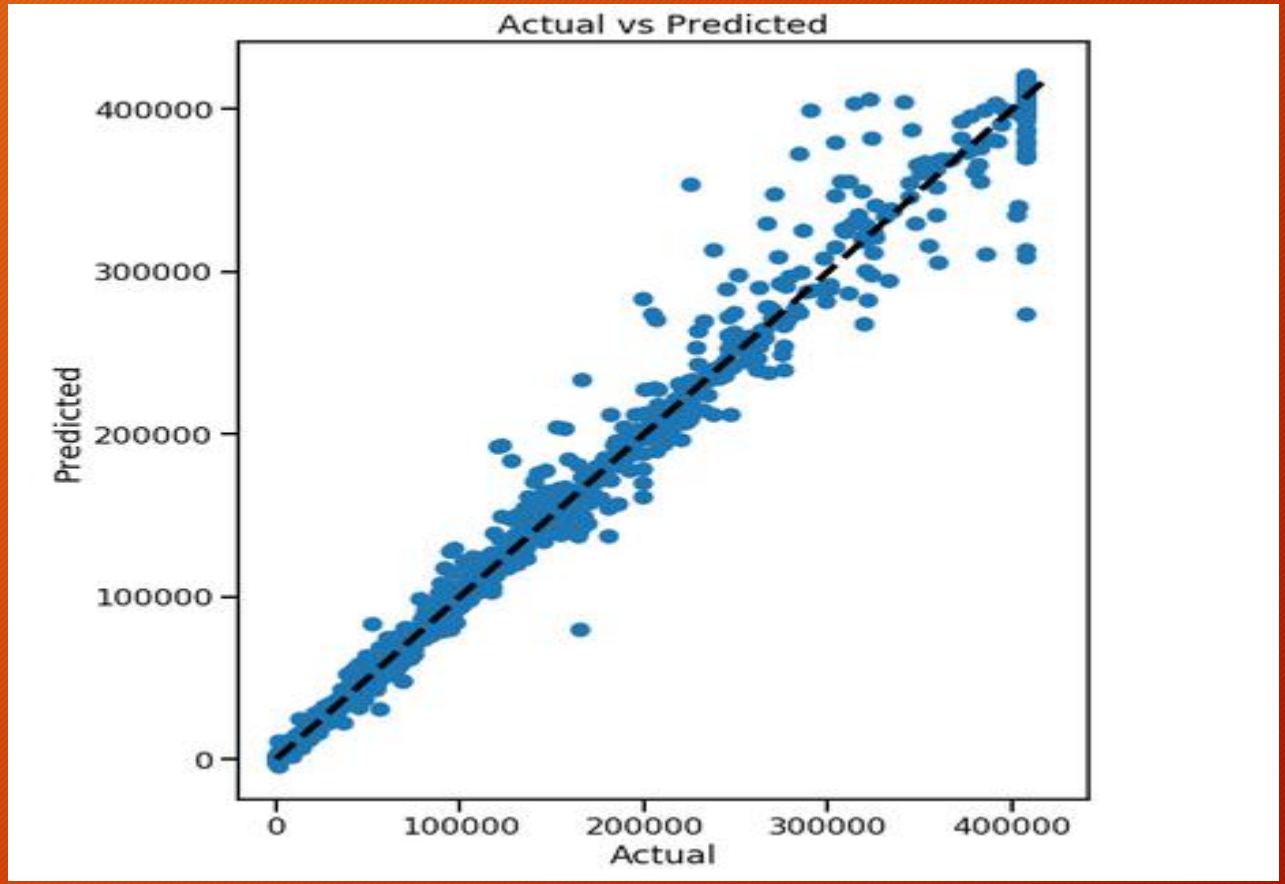
Out[437]:

	Train RMSE	Test RMSE	Training Score	Test Score
Gradient Boosting Regressor	17121.363784	18772.611970	0.984885	0.980767
LGBMRegressor	8838.888201	13918.219741	0.995972	0.989428
XGBRFRegressor	19531.705268	22176.163321	0.980330	0.973160

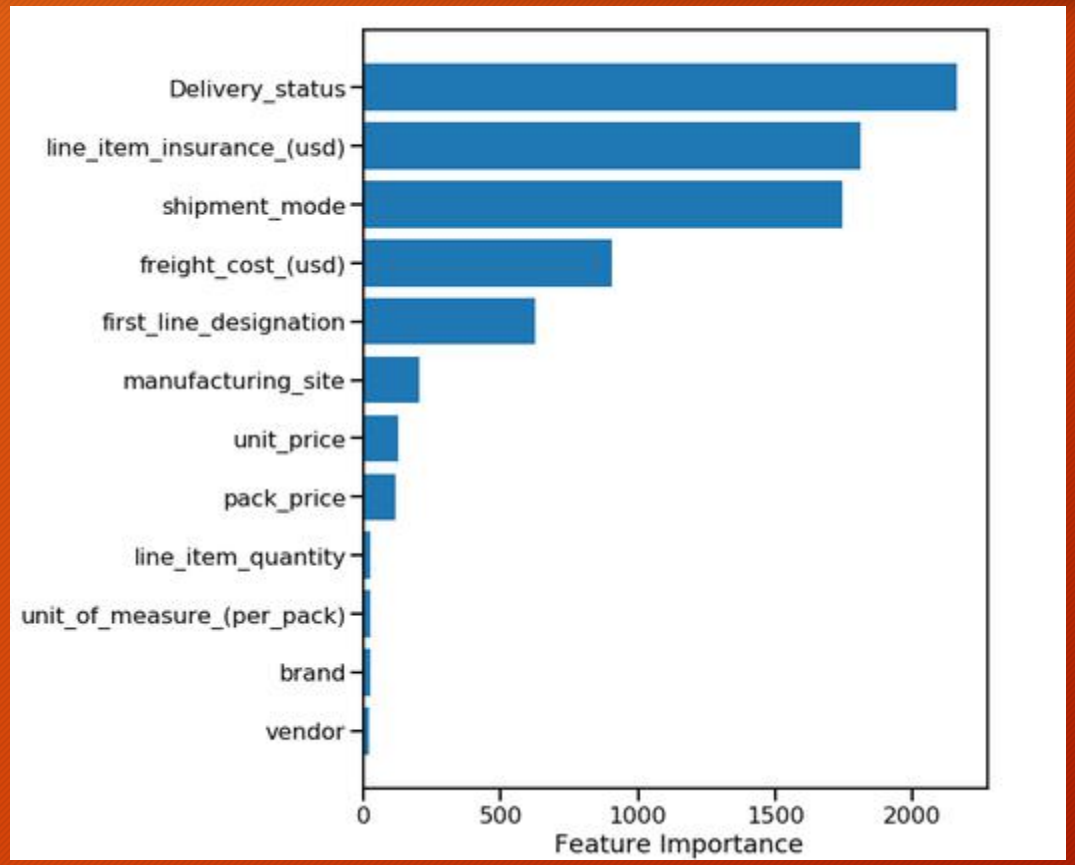
Out[438]:

	Train RMSE	Test RMSE	Training Score	Test Score
LGBMRegressor	6378.165869	12630.794511	0.997902	0.991293

Results


Actual vs Predicted
400000
300000
200000
100000
0
Predicted
0
100000
200000
300000
400000
Actual

Results



TOP 12 most important features

Model Performance Indicator

In [490]:
```
from sklearn.metrics import r2_score, mean_squared_error
y_train_pred =lgbm_tuned3.predict(x_train)
y_test_pred = lgbm_tuned3.predict(x_test)

print ("R-squared Training",r2_score(y_train, y_train_pred))
print ("R-squared Testing",r2_score(y_test, y_test_pred))
```

```
R-squared Training 0.9980930298933035
R-squared Testing 0.9920797654261521
```

In [500]:
```
#display adjusted R-squared train
1 - (1-0.998093)*(len(y_test)-1)/(len(y_test)-x_test.shape[1]-1)
```

Out[500]: 0.9980835265772479

In [501]:
```
#display adjusted R-squared test
1 - (1-0.99208)*(len(y_test)-1)/(len(y_test)-x_test.shape[1]-1)
```

Out[501]: 0.9920406557377048

Conclusion

- Dataset was cleaned , explored and visualized
- Top 12 correlating features were isolated
- Multiple Classical Machine Learning Algorithms and One Deep Learning based Neural Network was implemented against the data set to check the best performant algorithm.
- LightGBM ML Algorithm was selected for model traning and testing.
- Testing Accuracy is more than 99 precent.

THANK YOU