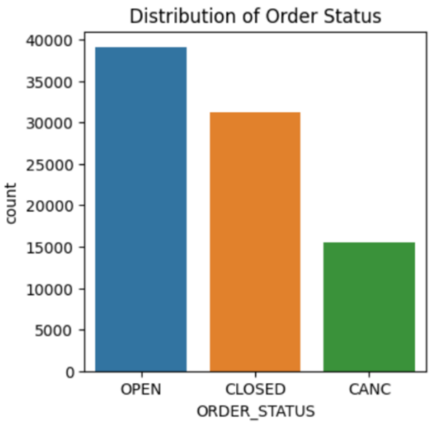
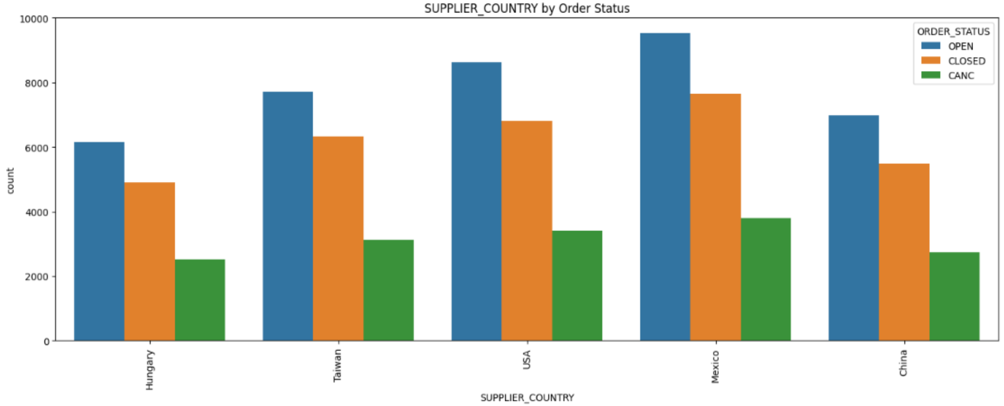
Procurement Purchase Prediction

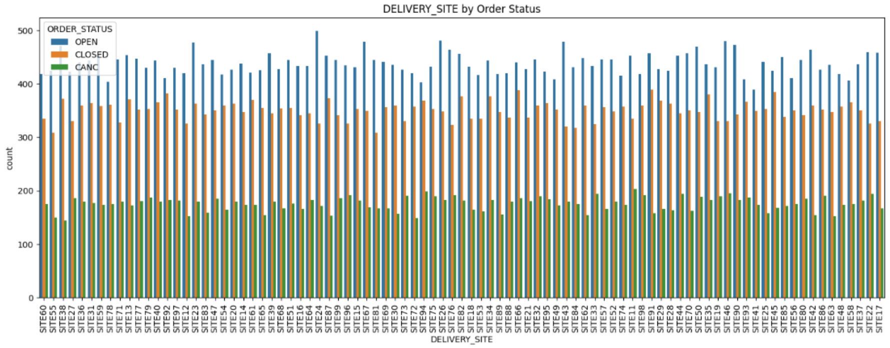
Business Use case	To identify if a purchase order would be cancelled or not
Presented by	Madhu

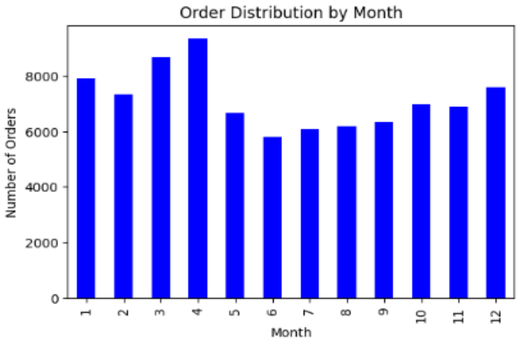
Meta data	RangeIndex: 85759 entries, 0 to 85758 Data columns (total 19 columns): ORDER_ID, ORDER_DATE, NEED_DATE, ORDER_STATUS, ORDER_CHANNEL, DELY_QTY, DELY_DATE, DELIVERY_SITE, BUYER_ID, BUYER_NAME, SUPPLIER_ID, SUPPLIER_NAME, SUPPLIER_COUNTRY, ITEM_ID, ORDER_QUANTITY, ORDER_UNIT_PRICE, ORDER_COST, ORDER_CURRENCY_CODE Dtypes: datetime64[ns](3), float64(3), int64(1), object(12)
Missing Values	DELY_QTY 54590 DELY_DATE 54590
General Observation	1. Target Variable: ORDER_STATUS (Canc/open/closed) 2. Whenever the order is closed delivery quantity and delivery date are available. 3. ORDER_ID is a unique identifier

Exploratory Data Analysis





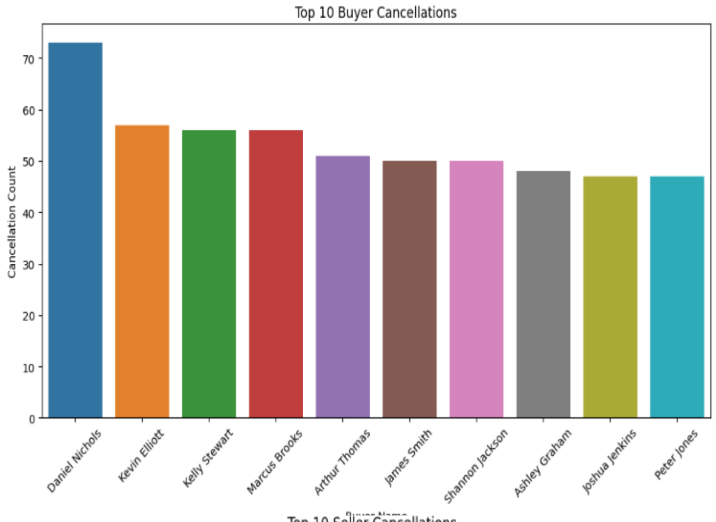


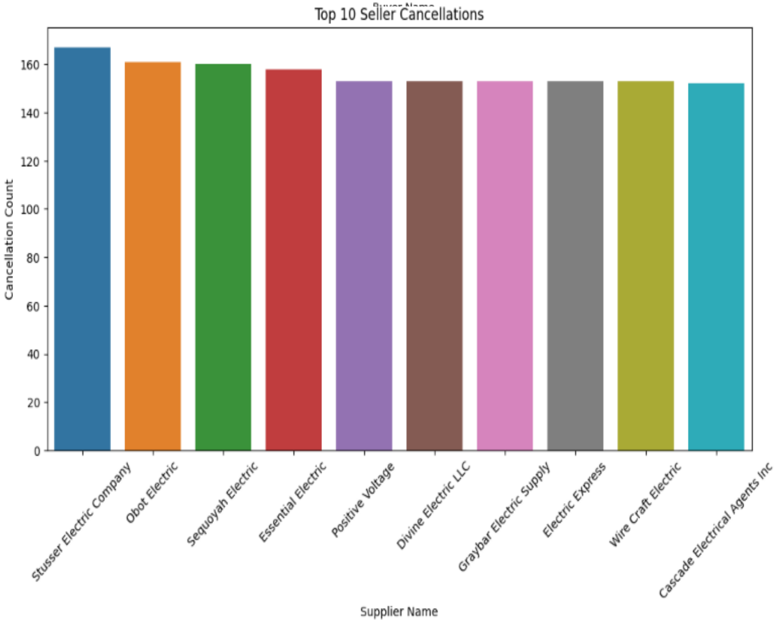


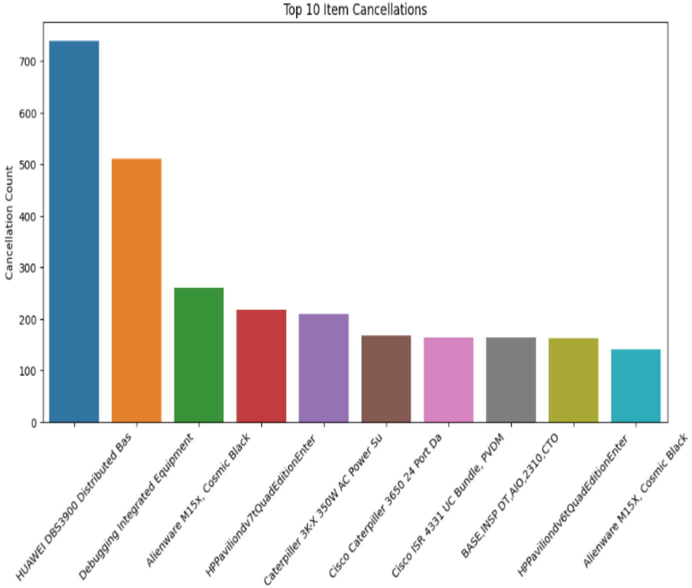
Exploratory Data Analysis- Buyer and Supplier

```
Top 5 Buyers based on total amount spent:
BUYER_NAME
Shannon Jackson     2.730083e+08
Daniel Nichols      2.566291e+08
Michael Callahan    1.964493e+08
Brian Green         1.919278e+08
Scott Bennett       1.911048e+08
Name: ORDER_COST, dtype: float64
```

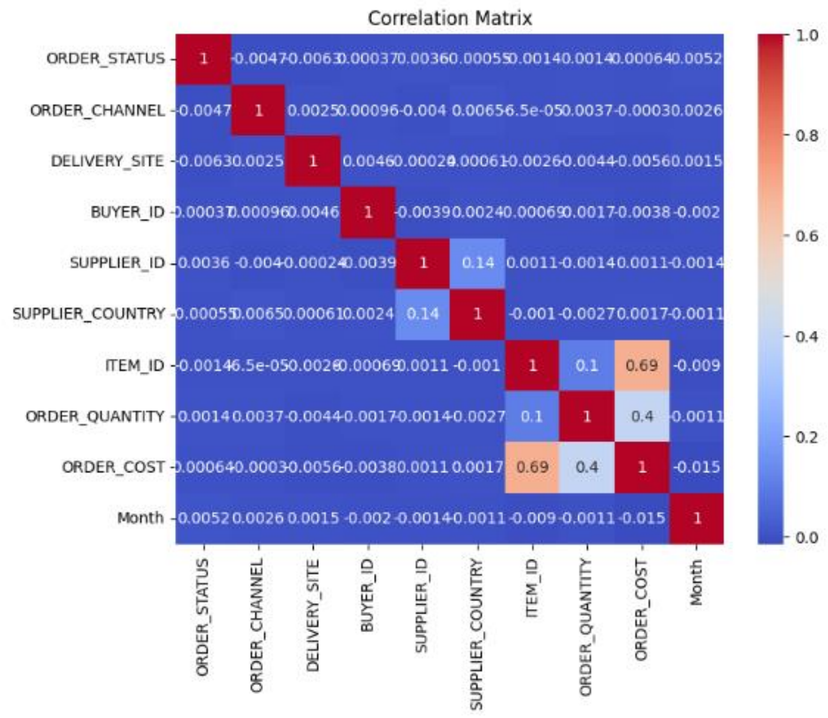
```
Top 5 Suppliers based on total amount received:
SUPPLIER_NAME
EC Electric                     6.470283e+08
Divine Electric LLC             6.173887e+08
Electric Express                6.170383e+08
NW Wind & Solar                 6.106914e+08
VECA Electric & Technologies    5.986096e+08
Name: ORDER_COST, dtype: float64
```







EDA - Feature Engineering



A correlation coefficient of 0.69 indicates a strong positive relationship between ITEM_ID and ORDER_COST. This means that as one variable increases, the other variable tends to increase as well.

```
Chi-square test of independence for DELIVERY_SITE:
Chi2 value: 173.65447662614497
P-value: 0.5358539431319053

Chi-square test of independence for SUPPLIER_COUNTRY:
Chi2 value: 6.677618438268236
P-value: 0.5717804797175376

Chi-square test of independence for ITEM_ID:
Chi2 value: 524.7806142430564
P-value: 0.7553159447592004

Chi-square test of independence for ORDER_STATUS:
Chi2 value: 171518.0
P-value: 0.0

Chi-square test of independence for ORDER_CHANNEL:
Chi2 value: 4.4797719858775755
P-value: 0.6120390995202432

Chi-square test of independence for BUYER_ID:
Chi2 value: 837.4504216664104
P-value: 0.9259634592102286

Chi-square test of independence for SUPPLIER_ID:
Chi2 value: 206.37796492029145
P-value: 0.8210821825390068
```

The interpretation of the chi-square test results : (Categorical Variables)
DELIVERY_SITE: Chi2 value: 173.65 P-value: 0.54 The chi-square test suggests that there is no significant association between the delivery site and the order status. The p-value is greater than the significance level of 0.05, indicating that the variables are independent.

SUPPLIER_COUNTRY: Chi2 value: 6.68 P-value: 0.57 The chi-square test suggests that there is no significant association between the supplier country and the order status. The p-value is greater than 0.05, indicating independence between the variables.

ITEM_ID: Chi2 value: 524.78 P-value: 0.76 The chi-square test suggests that there is no significant association between the item ID and the order status. The p-value is greater than 0.05, indicating independence.

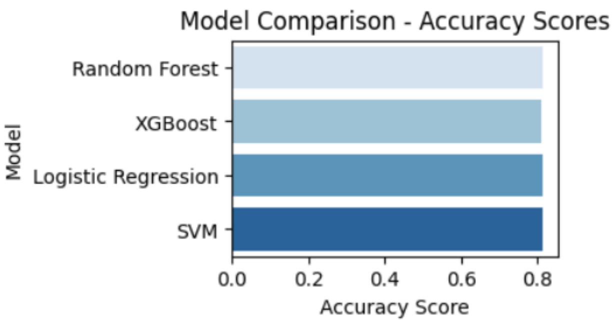
ORDER_CHANNEL: Chi2 value: 4.48 P-value: 0.61 The chi-square test suggests that there is no significant association between the order channel and the order status. The p-value is greater than 0.05, indicating independence.

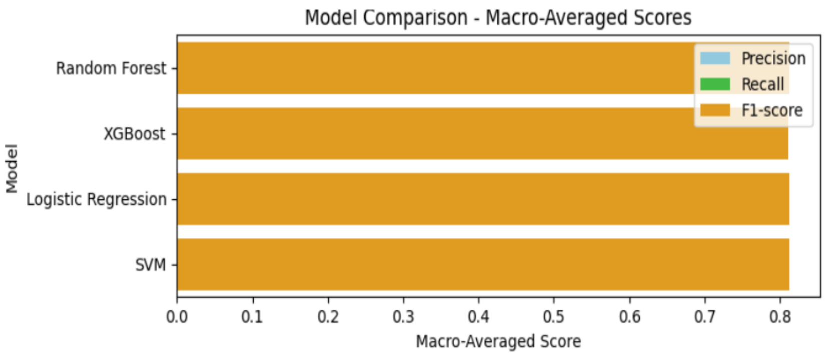
BUYER_ID: Chi2 value: 837.45 P-value: 0.93 The chi-square test suggests that there is no significant association between the buyer ID and the order status. The p-value is greater than 0.05, indicating independence.

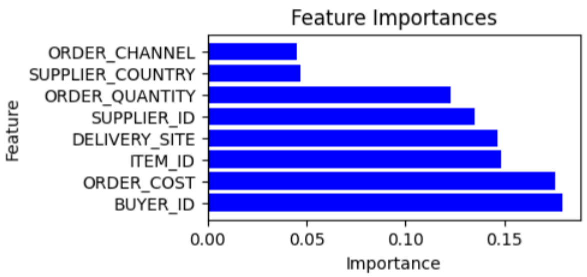
SUPPLIER_ID: Chi2 value: 206.38 P-value: 0.82 The chi-square test suggests that there is no significant association between the supplier ID and the order status. The p-value is greater than 0.05, indicating independence.

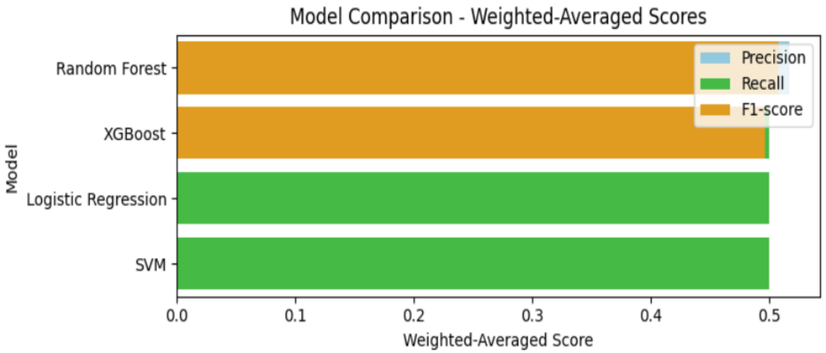
Model Insights and Analysis

Insignificant Features	**'ORDER_ID', 'NEED_DATE', 'DELY_DATE', ORDER_CURRENCY_CODE', 'ORDER_DATE', 'ORDER_UNIT_PRICE', 'ITEM_DESC', 'SUPPLIER_NAME', 'BUYER_NAME', 'DELY_QTY'**









- ✓ Experimented Random Forest, XGBoost, Logreg and SVM models.
- ✓ All the models achieves an overall accuracy of 0.81, indicating decent overall performance.
- ✓ However, when considering the performance across individual classes, the model's performance is relatively poor in terms of precision, recall, and F1-score (as indicated by macro-average).
- ✓ The weighted-average scores provide a more representative evaluation, accounting for class imbalance, and show better performance compared to the macro-average scores.
- ✓ Of these Random Forest has a better Macro avg.

Summary

EDA	✓ ✓	The p-value is greater than the significance level of 0.05, indicating that the variables DELIVERY_SITE, UPPLIER_COUNTRY, ITEM_ID, BUYER_ID, SUPPLIER_ID are independent. There is positive relationship between ITEM_ID and ORDER_COST
Feature Engineering	✓ ✓	Lable Encoding is performed for Categorical variables 'ORDER_CHANNEL', 'DELIVERY_SITE', 'BUYER_ID', 'SUPPLIER_ID', 'SUPPLIER_COUNTRY', 'ITEM_ID‘ Selection Bias - Insufficient data for he class : Cancel; To overcome the same, Oversamlping and downsampling is performed.
Models	✓ ✓ ✓ ✓ ✓ ✓ ✓	Removed Buyer name, Seller name and Item description as they can be interpreted using their IDs Since it is a classification problem with known target variable Supervised learning is performed (insights are explained in the previous slide) Stratified sampling is performed. As all the models have given an accuracy of 81%, it is understood that they have learnt similar features. F1 score of 90 signifies a strong classification performance and indicates that the models are effectively capturing the underlying patterns and relationships in the data. Have build a binary classification model combining Closed and Open as one class and retained Canc as the other class XGBoost is a powerful algorithm for handling imbalanced datasets, but the algorithm also has given an accuracy of 81%.
Problem	✓ ✓	Classification with class imbalance The class imbalance issue is challenging because the model is struggling to learn patterns from the minority class due to the dominance of the majority class.
Next Steps	✓ ✓ ✓ ✓ ✓	Explore other algorithms: Consider trying other machine learning algorithms that are suitable for imbalanced datasets, such as neural networks. Other feature engineering techniques specifically designed for addressing class imbalance, such as SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN (Adaptive Synthetic Sampling) can be studied More feature Engineering to create new features using domain knowledge. Hyperparameter Tuning like tuning max_depth, learning_rate, n_estimators etc Experiment ensemble such as bagging or boosting techniques like AdaBoost, CatBoost