Business Case

Products that do not sell over 20,000 units is could cause problems within Supply Chains due high production minimums.

Two Outcomes:

Significant Excess inventory or

New Product Does Not Get Produced

This problem is often **invisible** due to bias in the forecasting process.

Business Case

Build a statistical model that predicts if a product will sell above or below 20,000 units

Below 20k or Above 20k?

Logistic Regression Statistical Models

Models
1. Generalized Logistic Model (G.L.M.)
2. Decision Tree
3. Support Vector Machine
4. Random Forest
5. XGBoost

Confusion Matrix: Key Performance Indictors

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0
         1

               Accuracy
                  95% CI
    No Information Rate
    P-Value [Acc > NIR]

                  Kappa

 Mcnemar's Test P-Value

            Sensitivity
            Specificity
         Pos Pred Value
         Neg Pred Value
              Prevalence
         Detection Rate
   Detection Prevalence
      Balanced Accuracy

       'Positive' Class
```

Confusion Matrix

	0	1
0	True Positive	False Negative
1	False Positive	True Negative

Accuracy of Model

Sensitivity	**Correct Negative Predictions divided by the total number of Negatives**
Specificity	**Correct Positive Predictions divided by the total number of Positive**
Pos Pred	**Correct True Positive divided by the total number of True Positive and False Negative**
Neg Pred	**Correct True Negative divided by the total number of False Postive and True Negative**

1. Generalized Logistic Model (G.L.M.) Results

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  49 298
         1 360 856

               Accuracy : 0.579
                 95% CI : (0.5541, 0.6036)
    No Information Rate : 0.7383
    P-Value [Acc > NIR] : 1.00000

                  Kappa : -0.1455

 Mcnemar's Test P-Value : 0.01741

            Sensitivity : 0.11980
            Specificity : 0.74177
         Pos Pred Value : 0.14121
         Neg Pred Value : 0.70395
             Prevalence : 0.26168
         Detection Rate : 0.03135
   Detection Prevalence : 0.22201
      Balanced Accuracy : 0.43079

       'Positive' Class : 0
```

2. Decision Tree Model Results

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  110  237
         1   74 1142

               Accuracy : 0.801
                 95% CI : (0.7804, 0.8206)
    No Information Rate : 0.8823
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3078

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.59783
            Specificity : 0.82814
         Pos Pred Value : 0.31700
         Neg Pred Value : 0.93914
             Prevalence : 0.11772
         Detection Rate : 0.07038
   Detection Prevalence : 0.22201
      Balanced Accuracy : 0.71298

       'Positive' Class : 0
```

3. Support Vector Machine Model Results

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  43 166
         1  10 878

               Accuracy : 0.8396
                 95% CI : (0.8165, 0.8608)
    No Information Rate : 0.9517
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2721

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.81132
            Specificity : 0.84100
         Pos Pred Value : 0.20574
         Neg Pred Value : 0.98874
             Prevalence : 0.04831
         Detection Rate : 0.03920
   Detection Prevalence : 0.19052
      Balanced Accuracy : 0.82616

       'Positive' Class : 0
```

4. Random Forest Model Results

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  95 114
         1  44 844

               Accuracy : 0.856
                 95% CI : (0.8338, 0.8762)
    No Information Rate : 0.8733
    P-Value [Acc > NIR] : 0.9597

                  Kappa : 0.4645

 Mcnemar's Test P-Value : 4.034e-08

            Sensitivity : 0.6835
            Specificity : 0.8810
         Pos Pred Value : 0.4545
         Neg Pred Value : 0.9505
             Prevalence : 0.1267
         Detection Rate : 0.0866
   Detection Prevalence : 0.1905
      Balanced Accuracy : 0.7822

       'Positive' Class : 0
```

5. XGBoost Model Results

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  81 128
         1  38 850

               Accuracy : 0.8487
                 95% CI : (0.8261, 0.8694)
    No Information Rate : 0.8915
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4127

 Mcnemar's Test P-Value : 4.924e-12

            Sensitivity : 0.68067
            Specificity : 0.86912
         Pos Pred Value : 0.38756
         Neg Pred Value : 0.95721
             Prevalence : 0.10848
         Detection Rate : 0.07384
   Detection Prevalence : 0.19052
      Balanced Accuracy : 0.77490

       'Positive' Class : 0
```

Model Comparison

Statistical Model	Accuracy	Sensitivity	Specificity	Pos Pred	Neg Pred
Generalized Logistic Model	60%	8%	77%	10%	72%
Decision Tree	84%	63%	86%	34%	95%
Support Vector Machine	83%	81%	84%	21%	99%
Random Forest	86%	68%	88%	46%	95%
XGBoost	85%	68%	87%	39%	96%