LA SALLE
CAMPUS
BARCELONA

STARTUPS PROJECT

CAN WE PREDICT STARTUP SUCCESS USING TWITTER?

DATA MINING

METHODOLOGY

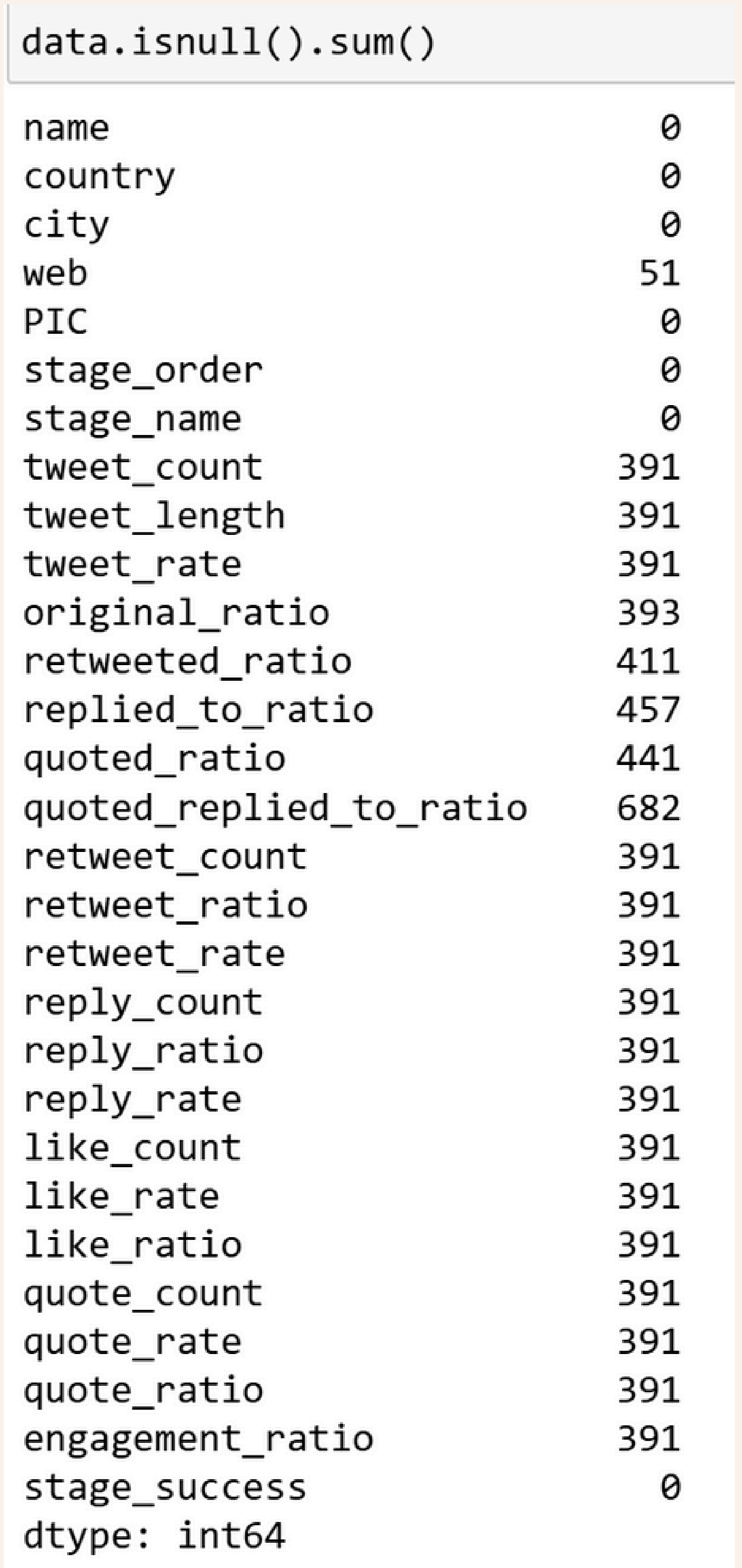
The primary goal of this project is to **predict the success of startups** in transitioning from the early stages of ideation and **inception to the scale-up** stage by analyzing their Twitter activity. By leveraging **social media analytics,** we aim to identify key indicators of startup potential and success.

In the digital age, social media platforms like Twitter (X) have become critical tools for startups to engage with their audience, build their brand, and attract investors.

Twitter activity can provide valuable insights into a startup's market presence, customer engagement, and overall momentum.

This project seeks to harness these insights to forecast which startups are likely to progress to the next stage of development.

DATA PREPARATION



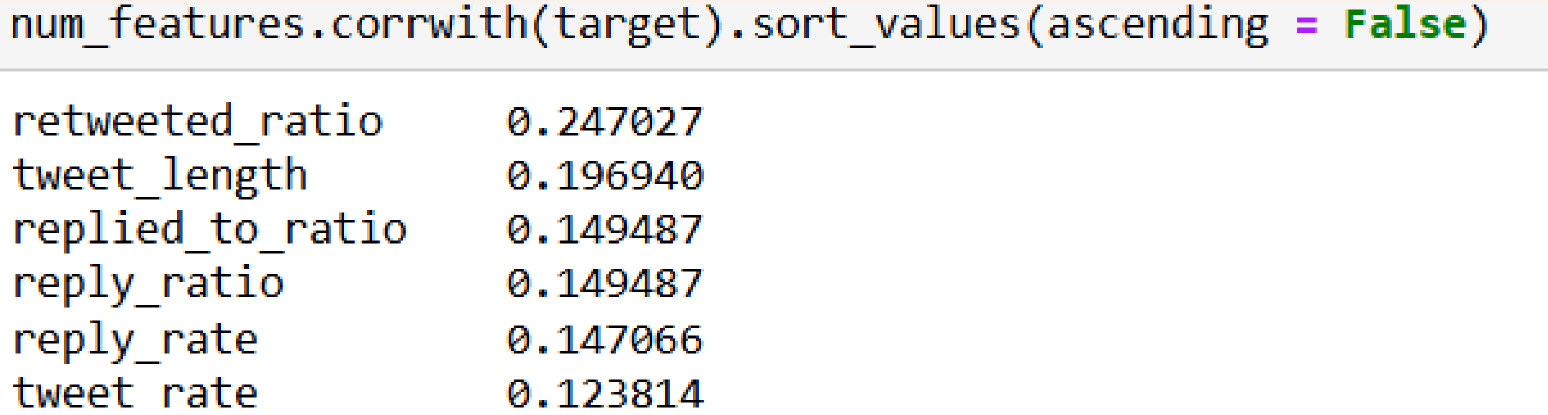
```
data.isnull().sum()
```

```
name                        0
country                     0
city                        0
web                        51
PIC                         0
stage_order                 0
stage_name                  0
tweet_count               391
tweet_length              391
tweet_rate                391
original_ratio            393
retweeted_ratio           411
replied_to_ratio          457
quoted_ratio              441
quoted_replied_to_ratio   682
retweet_count             391
retweet_ratio             391
retweet_rate              391
reply_count               391
reply_ratio               391
reply_rate                391
like_count                391
like_rate                 391
like_ratio                391
quote_count               391
quote_rate                391
quote_ratio               391
engagement_ratio          391
stage_success               0
dtype: int64
```

DROPPED COLUMNS

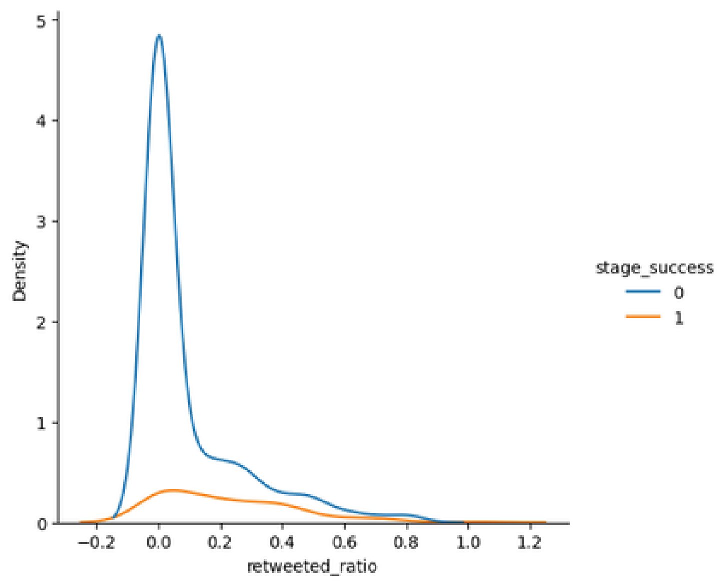
quoted_replied_to_ratio
WEB
PIC
Stage_Name
Stage_Number

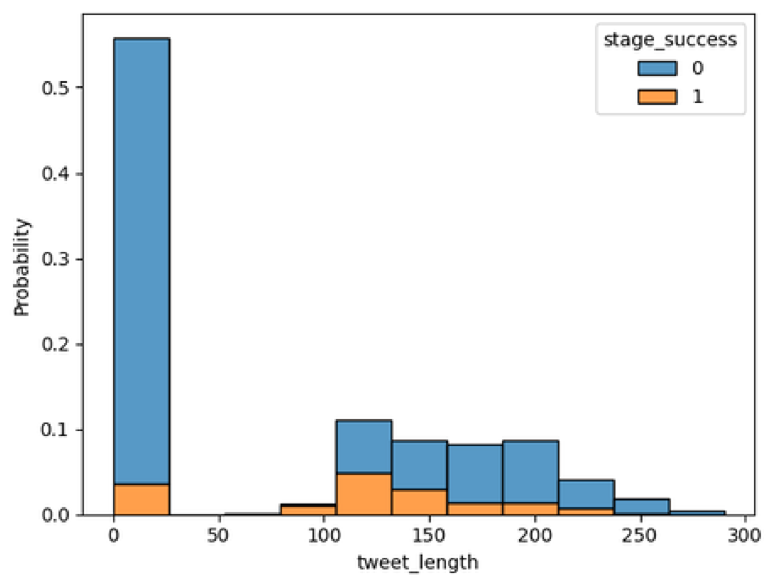
DATA EXPLORATION



```
num_features.corrwith(target).sort_values(ascending = False)
```

```
retweeted_ratio     0.247027
tweet_length        0.196940
replied_to_ratio    0.149487
reply_ratio         0.149487
reply_rate          0.147066
tweet_rate          0.123814
```



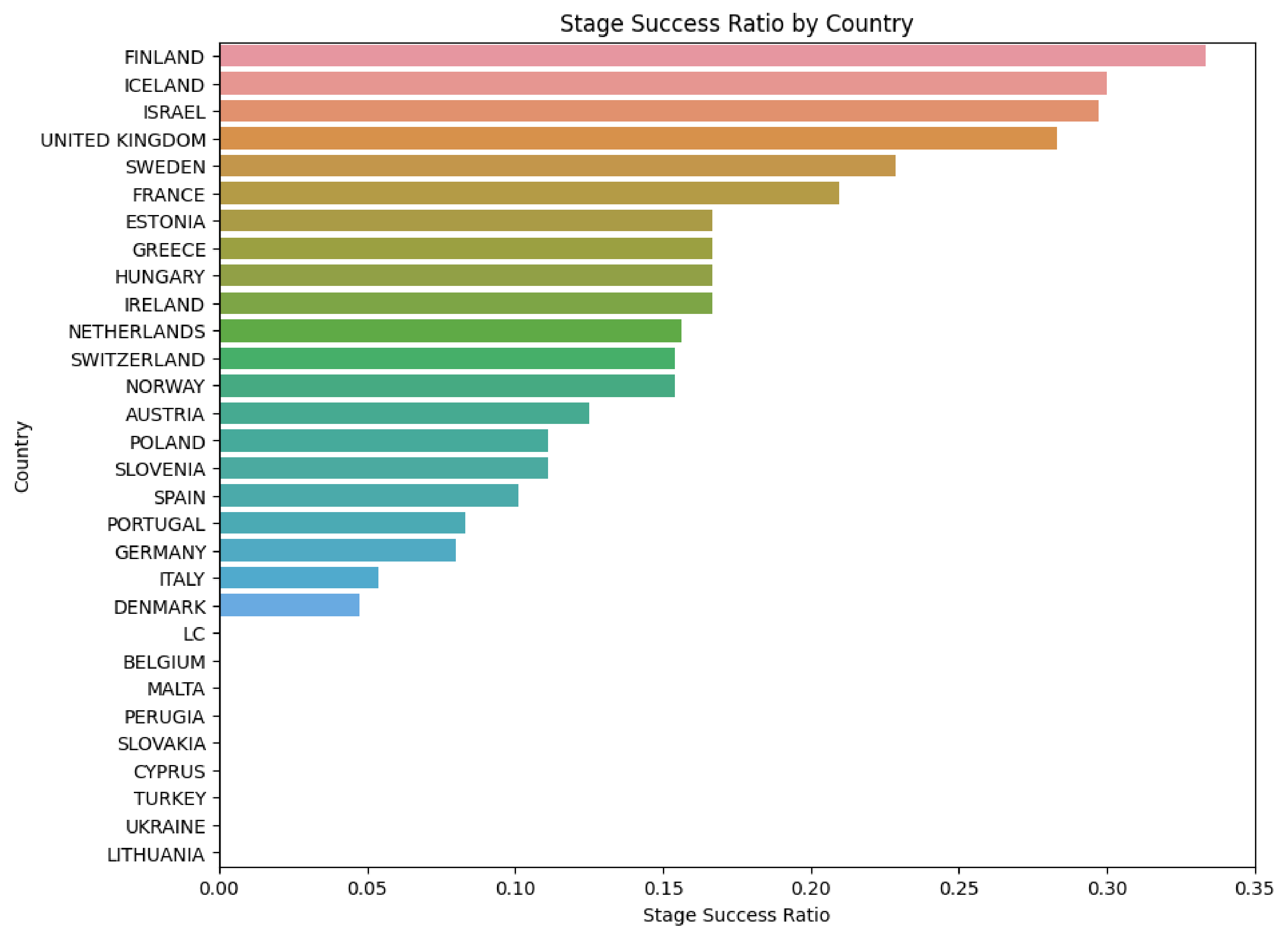


- **retweeted_ratio, tweet_length, and replied_to_ratio** demonstrated relatively **strong positive correlations** with the target variable. This suggests that posts with higher retweet ratios, longer lengths, and higher replied-to ratios tend to have higher levels of engagement, as indicated by the target variable.

- **original_ratio, retweet_count, and like_count** exhibited negative correlations with the target variable, suggesting a decrease in engagement.

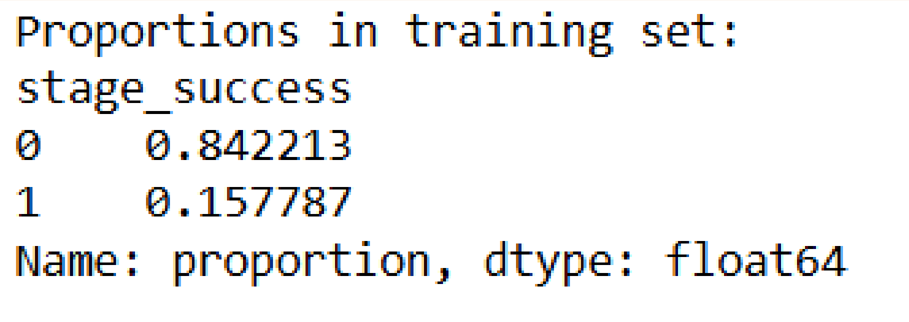
- Short tweet length might indicate concise and impactful messaging and **tweets between 100-150 characters** strike a balance between sufficient information and maintaining audience engagement.

- **Finland** seems to have the highest stage success ratio, followed closely by **Iceland, Israel and the United Kingdom.**

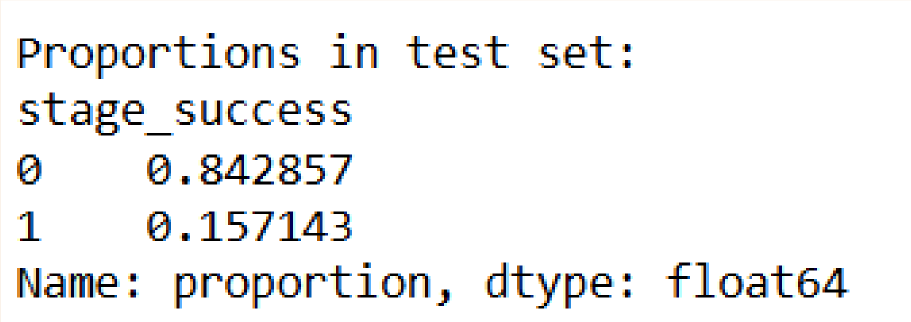


BUILD AND FIT THE MODEL

01 Making sure we have the same proportions of data in each sets:

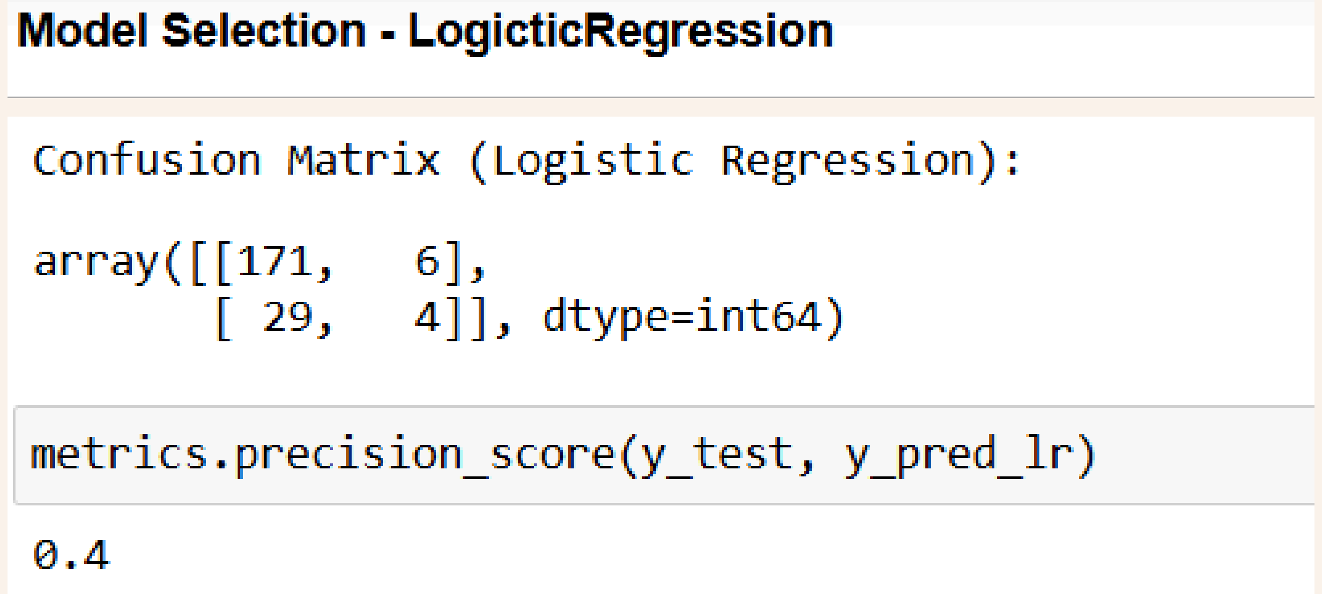


```
Proportions in training set:
stage_success
0    0.842213
1    0.157787
Name: proportion, dtype: float64
```



```
Proportions in test set:
stage_success
0    0.842857
1    0.157143
Name: proportion, dtype: float64
```

02 Choosing the model:



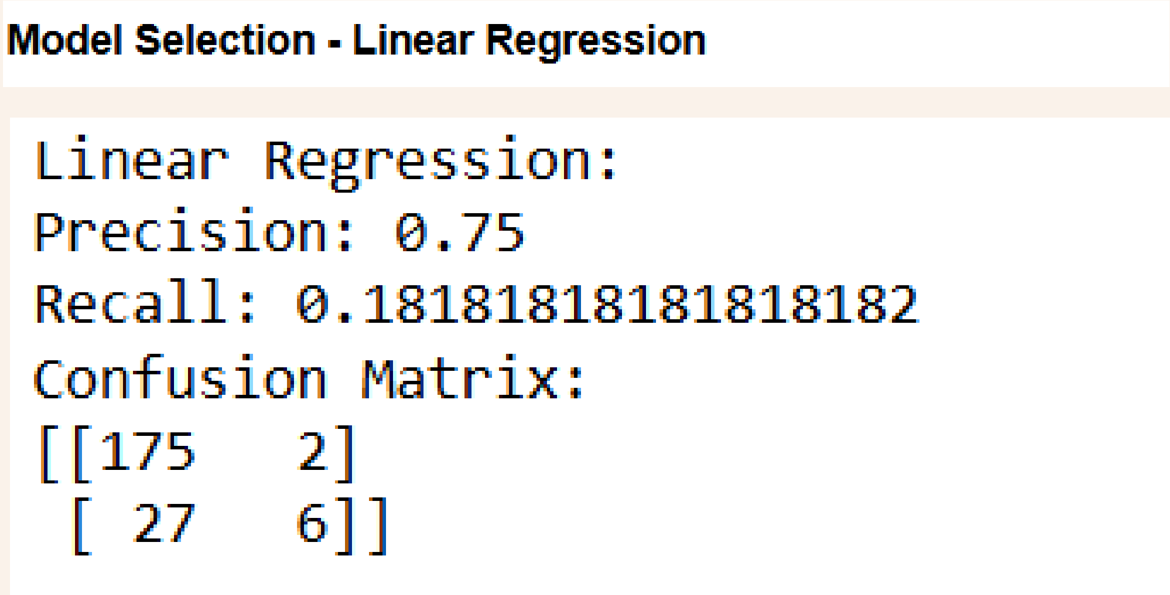
Model Selection - LogicticRegression

```
Confusion Matrix (Logistic Regression):

array([[171,   6],
       [ 29,   4]], dtype=int64)
```

```
metrics.precision_score(y_test, y_pred_lr)
```

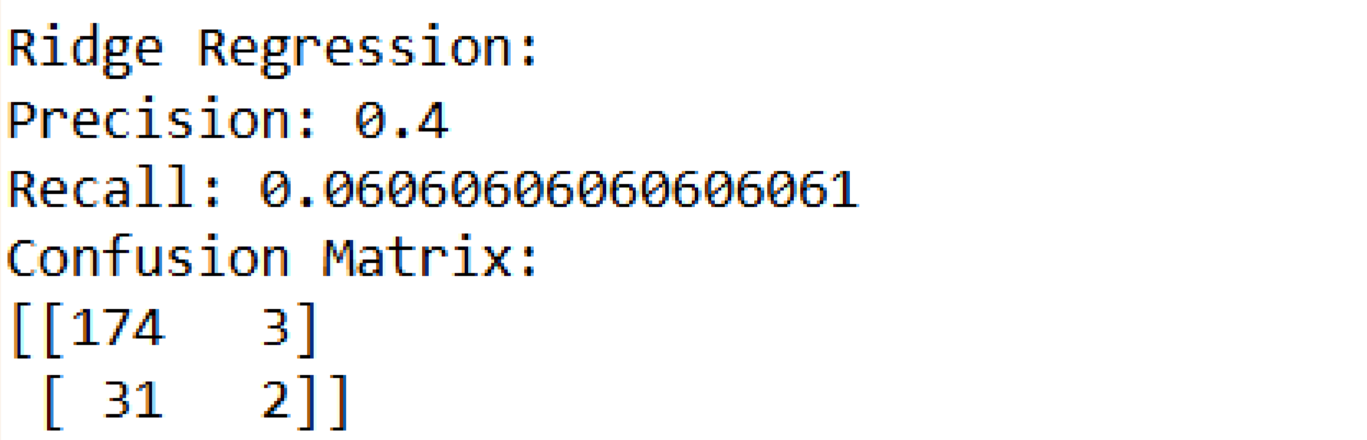
```
0.4
```



Model Selection - Linear Regression

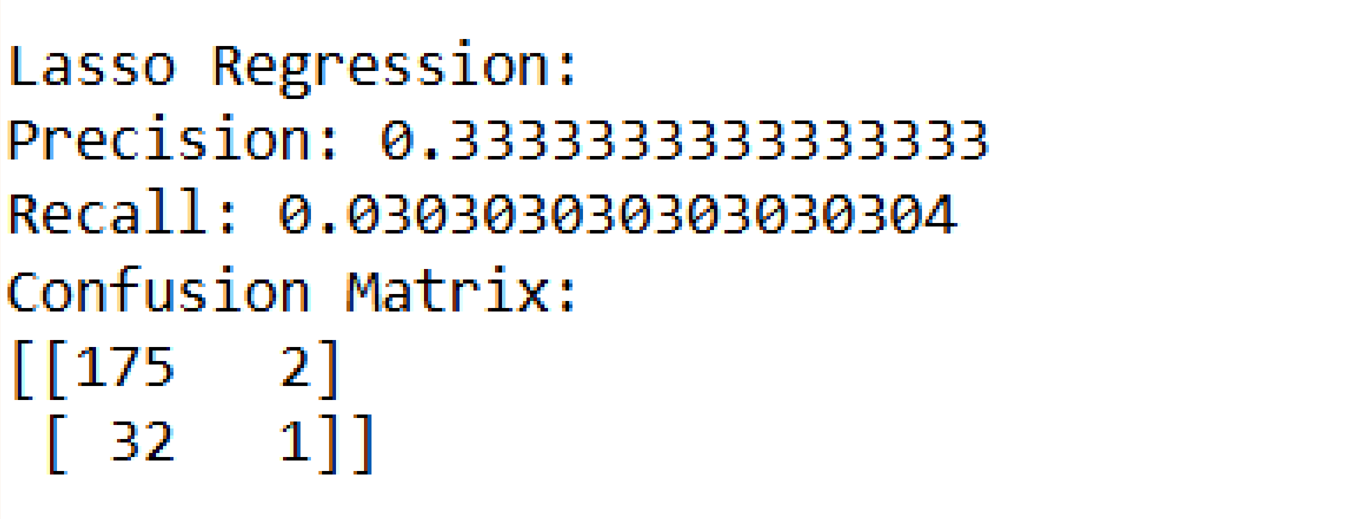
```
Linear Regression:
Precision: 0.75
Recall: 0.18181818181818182
Confusion Matrix:
[[175   2]
 [ 27   6]]
```

Model Selection - Ridge Regression



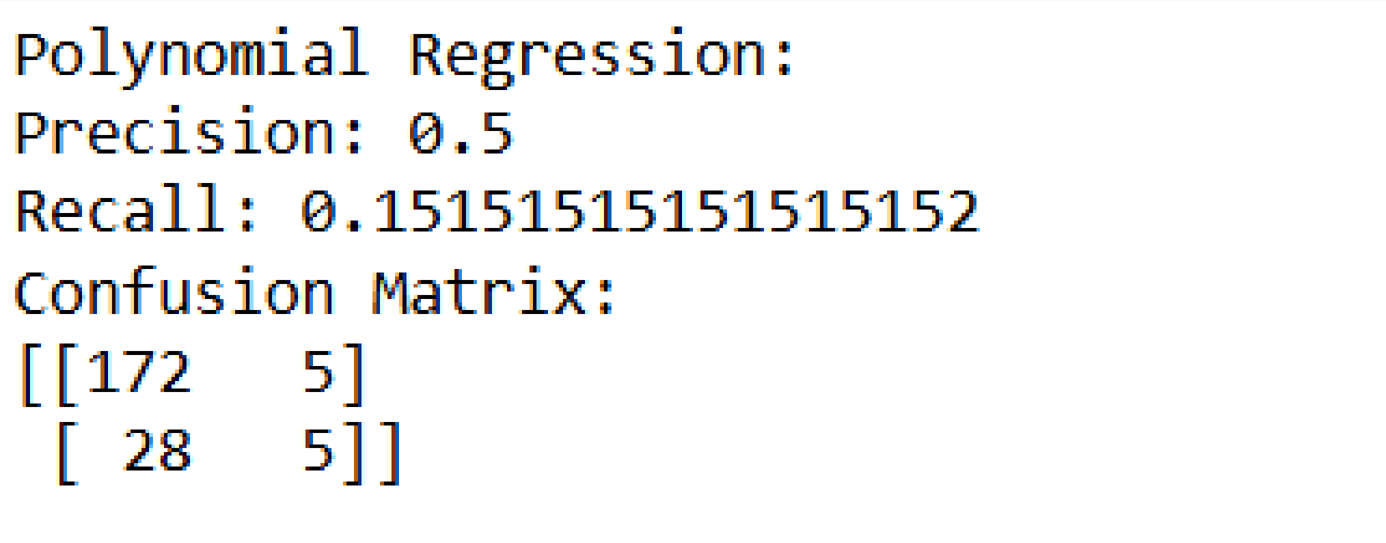
```
Ridge Regression:
Precision: 0.4
Recall: 0.06060606060606061
Confusion Matrix:
[[174   3]
 [ 31   2]]
```

Model Selection - Lasso Regression



```
Lasso Regression:
Precision: 0.3333333333333333
Recall: 0.030303030303030304
Confusion Matrix:
[[175   2]
 [ 32   1]]
```

Model Selection - Polynomial Regression



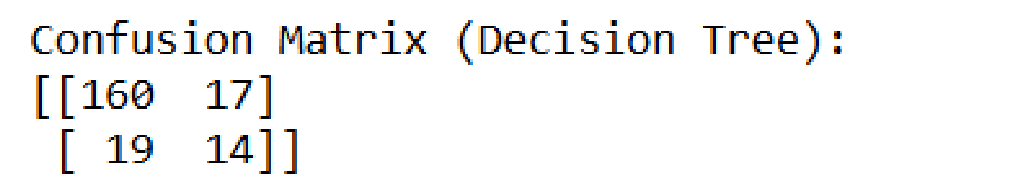
```
Polynomial Regression:
Precision: 0.5
Recall: 0.15151515151515152
Confusion Matrix:
[[172   5]
 [ 28   5]]
```

Evaluation: The precision and recall scores are quite low for each of these models. The confusion matrixes indicate misclassification which suggests it would be better to try more complex models.

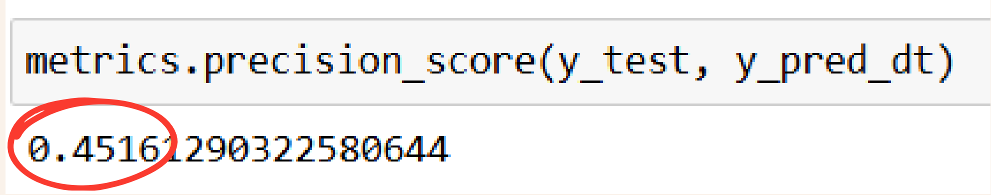
Decision: We reject these models due to their low precision and recall scores, indicating poor predictive performance.

Second Model: DecisionTreeClassifier

- **Without any hyperparameters:**

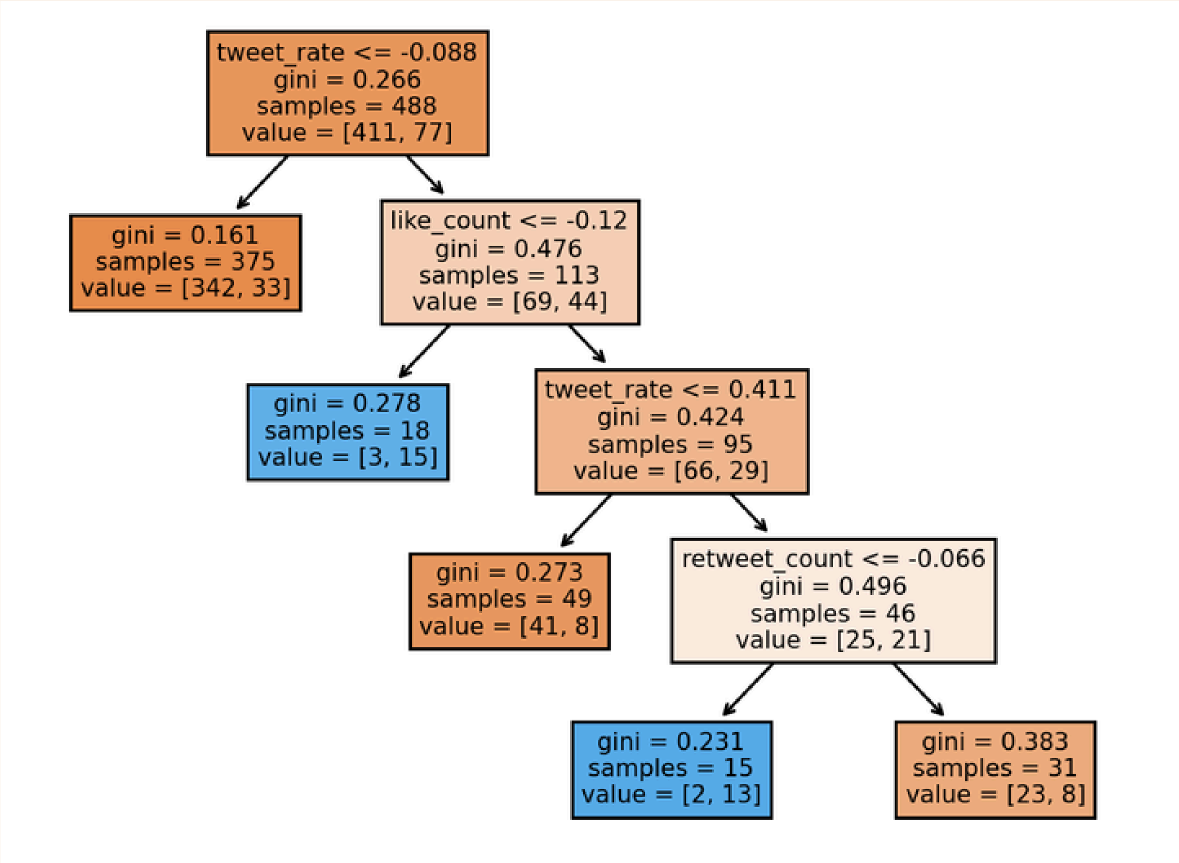


```
Confusion Matrix (Decision Tree):
[[160  17]
 [ 19  14]]
```

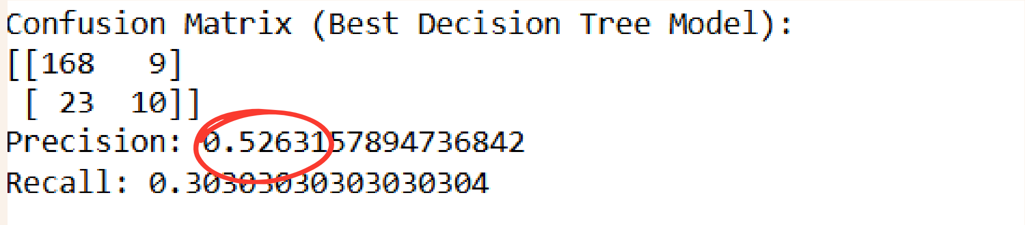


```
metrics.precision_score(y_test, y_pred_dt)
0.45161290322580644
```

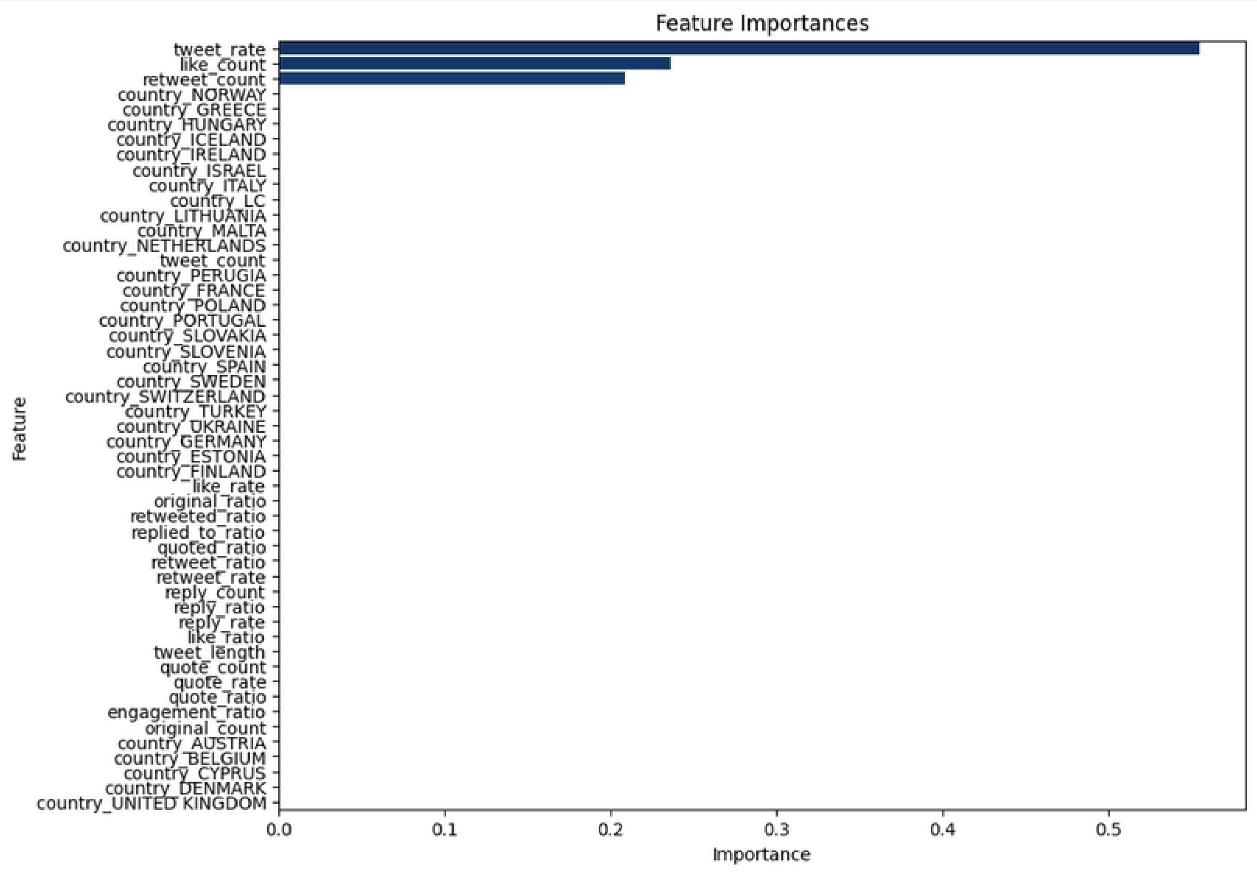
Decision: Although the precision score improved slightly, the recall score is still relatively low, and we're not yet achieving our target of predicting at least 5 successful companies with high precision.



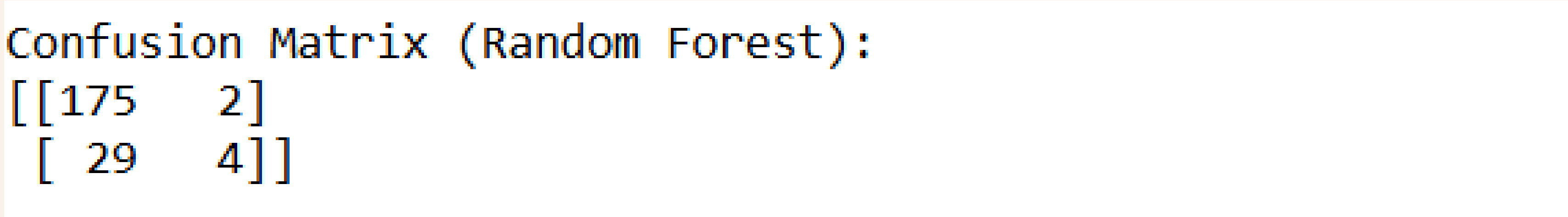
- **With hyperparameters:**



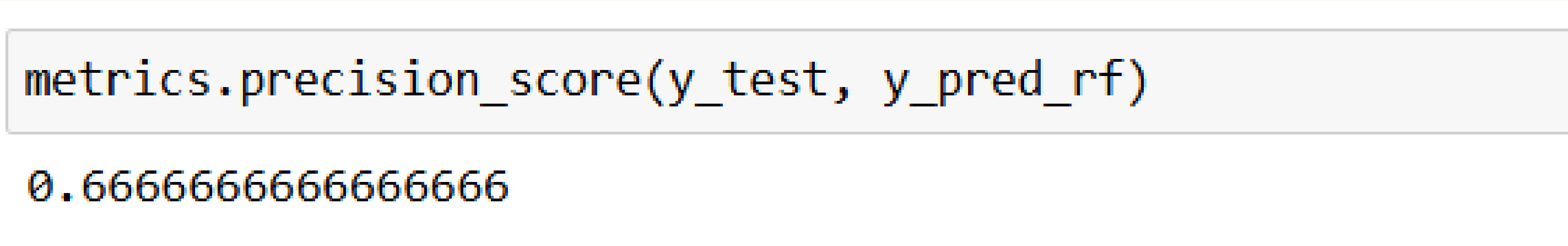
```
Confusion Matrix (Best Decision Tree Model):
[[168   9]
 [ 23  10]]
Precision: 0.5263157894736842
Recall: 0.30303030303030304
```



Third Model: RandomForestClassifier



```
Confusion Matrix (Random Forest):
[[175   2]
 [ 29   4]]
```

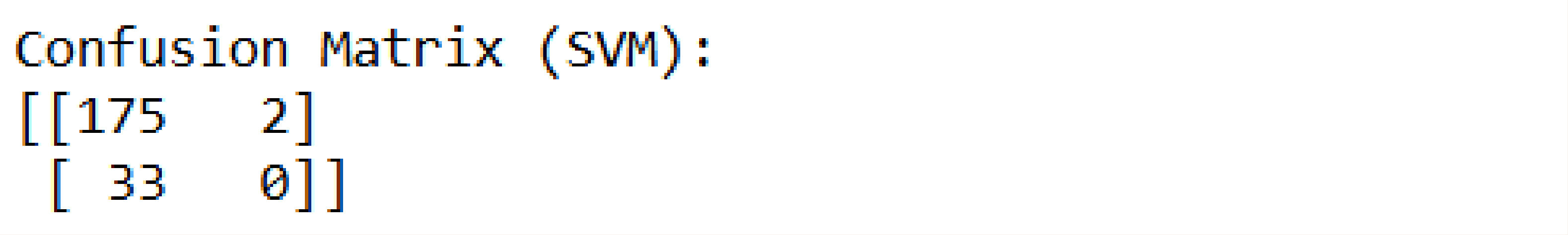


```
metrics.precision_score(y_test, y_pred_rf)
```

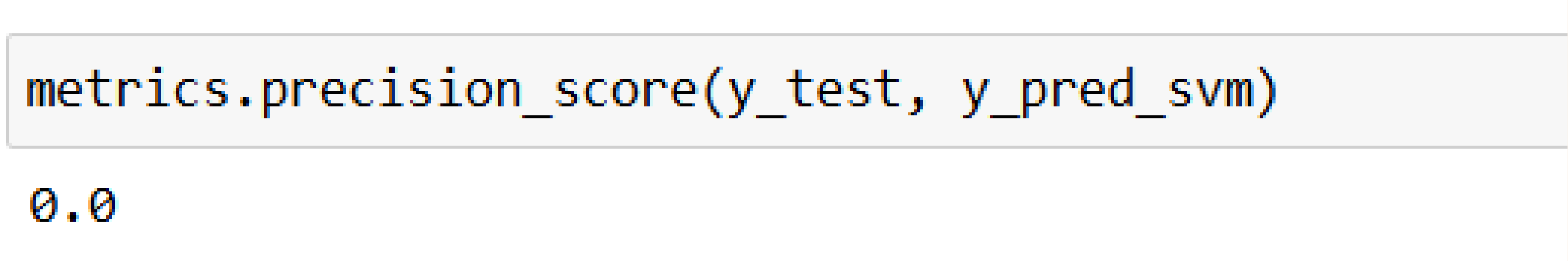
```
0.6666666666666666
```

Decision: The RandomForestClassifier achieved a relatively high precision score of 0.67, which meets our criterion. However, the recall score is still relatively low, indicating that it might not be predicting enough successful companies. While the precision is acceptable, the model doesn't meet our target of predicting at least 5 successful companies with high precision.

Forth Model: Support Vector Machines



```
Confusion Matrix (SVM):
[[175   2]
 [ 33   0]]
```

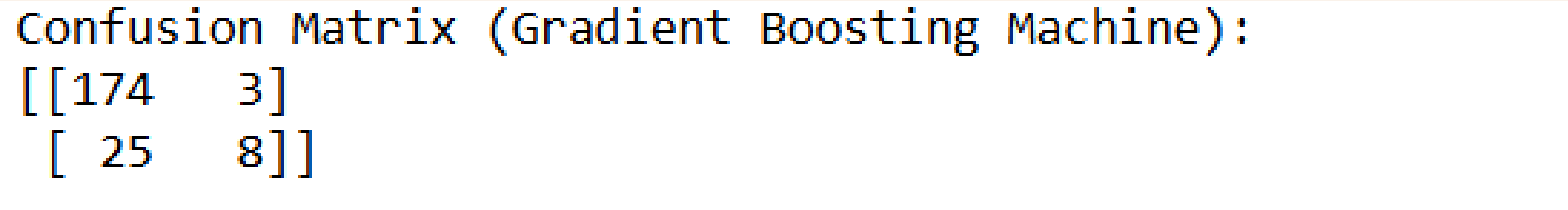


```
metrics.precision_score(y_test, y_pred_svm)
```

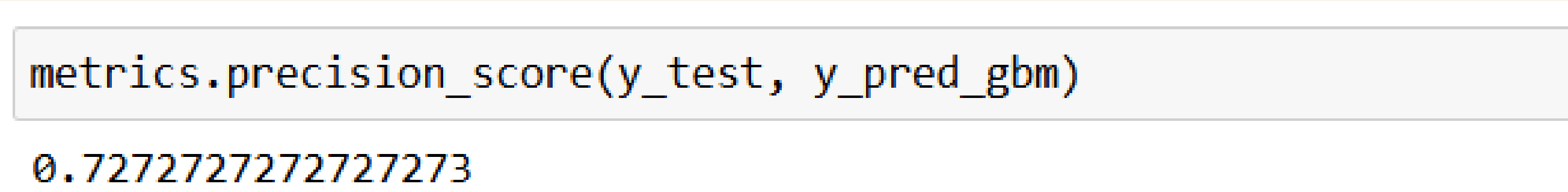
```
0.0
```

Decision: Surprisingly, the SVM model predicted all companies as unsuccessful (0) in the test set. This resulted in both precision and recall scores of 0, indicating that the model didn't predict any successful companies. This makes the SVM model unsuitable for our task of predicting successful companies with high precision.

Fifth Model: GradientBoostingClassifier



```
Confusion Matrix (Gradient Boosting Machine):
[[174   3]
 [ 25   8]]
```

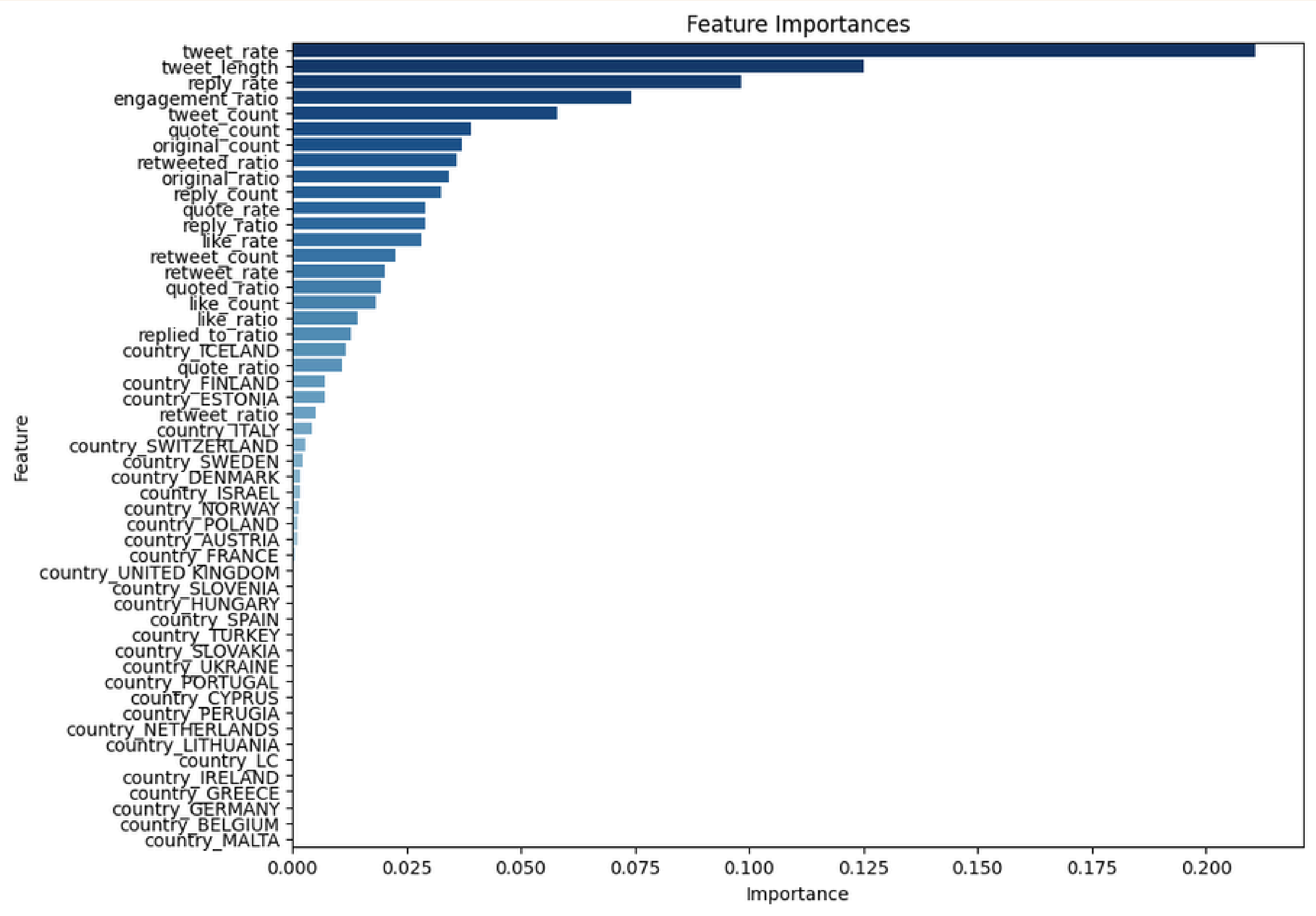


```
metrics.precision_score(y_test, y_pred_gbm)
```

```
0.7272727272727273
```

Despite attempts to optimize our Gradient Boosting Classifier's performance by adjusting hyperparameters, including **learning_rate and max_depth**, we consistently observed decreased precision and increased false positives.

Given the default model's already satisfactory precision score of 0.7273, we chose to maintain these settings.



Top 5 Companies

Company	Country	Probability
INDOORATLAS OY	Finland	0.9580
UNBABEL, LDA	Portugal	0.9020
BIOSERVO TECHNOLOGIES AB	Sweden	0.8415
ILLUSIVE NETWORKS LTD	Isreal	0.8285
AIMOTIVE INFORMATIKAI KORLATOLT FELELOSSEGU TARSASAG	Hungary	0.7835

FINAL PROFIT

Investment

You allocated 1 million dollars to each of the five companies, totaling 5 million dollars in investments.

Outcome

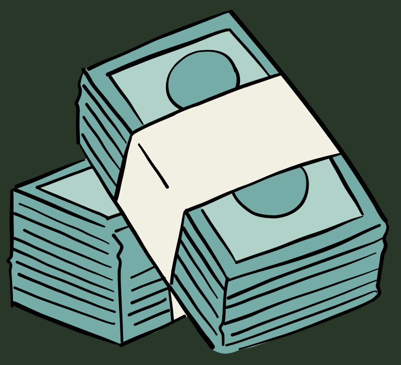
- True Positives (TP): 8 companies performed well as predicted.
- False Positives (FP): 3 companies were predicted to perform well but didn't.

Total Benefit

- Successful Investments (TP): $14,545,454.6
- Unsuccessful Investments (FP): -$272,727.272

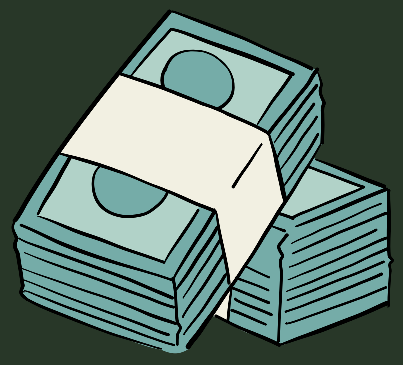
Benefit Calculation

- Benefit = Successful Investments - Unsuccessful Investments
- Benefit = $14,272,727.3



Final Profit

Final profit stands at **$9,272,727.3** after deducting the initial investment.



CONCLUSION

Our investment approach, guided by predictive modeling, generated an impressive profit of approximately **9.27 million dollars.**

By strategically allocating investments **based on probability scores**, we identified successful companies while minimizing losses.

This highlights the effectiveness of our model, particularly **Gradient Boosting**, in informing investment decisions.

LA SALLE
CAMPUS
BARCELOAN

DATA MINING

THANK YOU

Questions?