Bull or Bear: Stock Market Predictions

S&P 500 price development for tomorrow

David Thrien, 28. March 2025

Content

1. Target and Data Basis
2. Cleaning and Feature Engineering
3. Logistic Regression and Sampling Manipulation
4. Random Forest Classifier
5. Summary

Target and Data Basis

- **Target:**
  - Overview about technical analysis indicators and their impact on tomorrow’s price predictions
  - Development of a model to predict if the price of the S&P 500 will increase or decrease during the next day (open to close)

- **Data Basis**
  - S&P 500 E-mini Futures prices from April 2012 until end of March 2025 (12 years) on a daily timeframe
  - 10 price indicators (trend + oscillators) with additional internal indicators and volume

Target and Data Basis

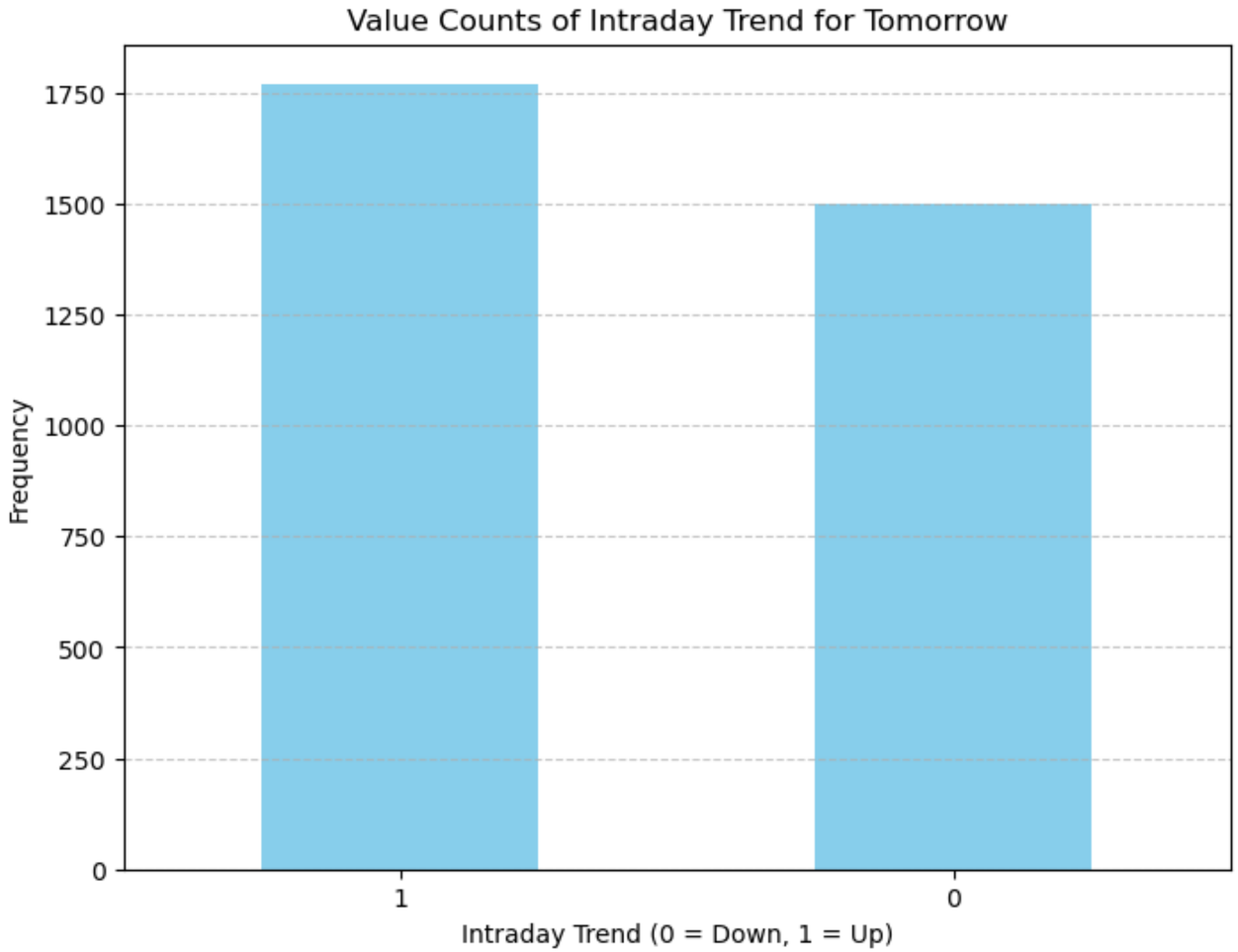
- **Why to go for categorical instead of numerical prediction:**
  - Technical indicators are specialized on direction instead of valuation
  - A 1 Day prediction with daily data is very short term and usually has a lot of noise

- **Quick Check with Lazy Regressor:**
  - R2 = 1, as there is a very high correlation between most indicators and price
  - Best model has an RMSE of 32.7 while the mean absolute difference between the close of today and the close of tomorrow is 21.71 —> no value

Cleaning and Feature Engineering

Target Creation

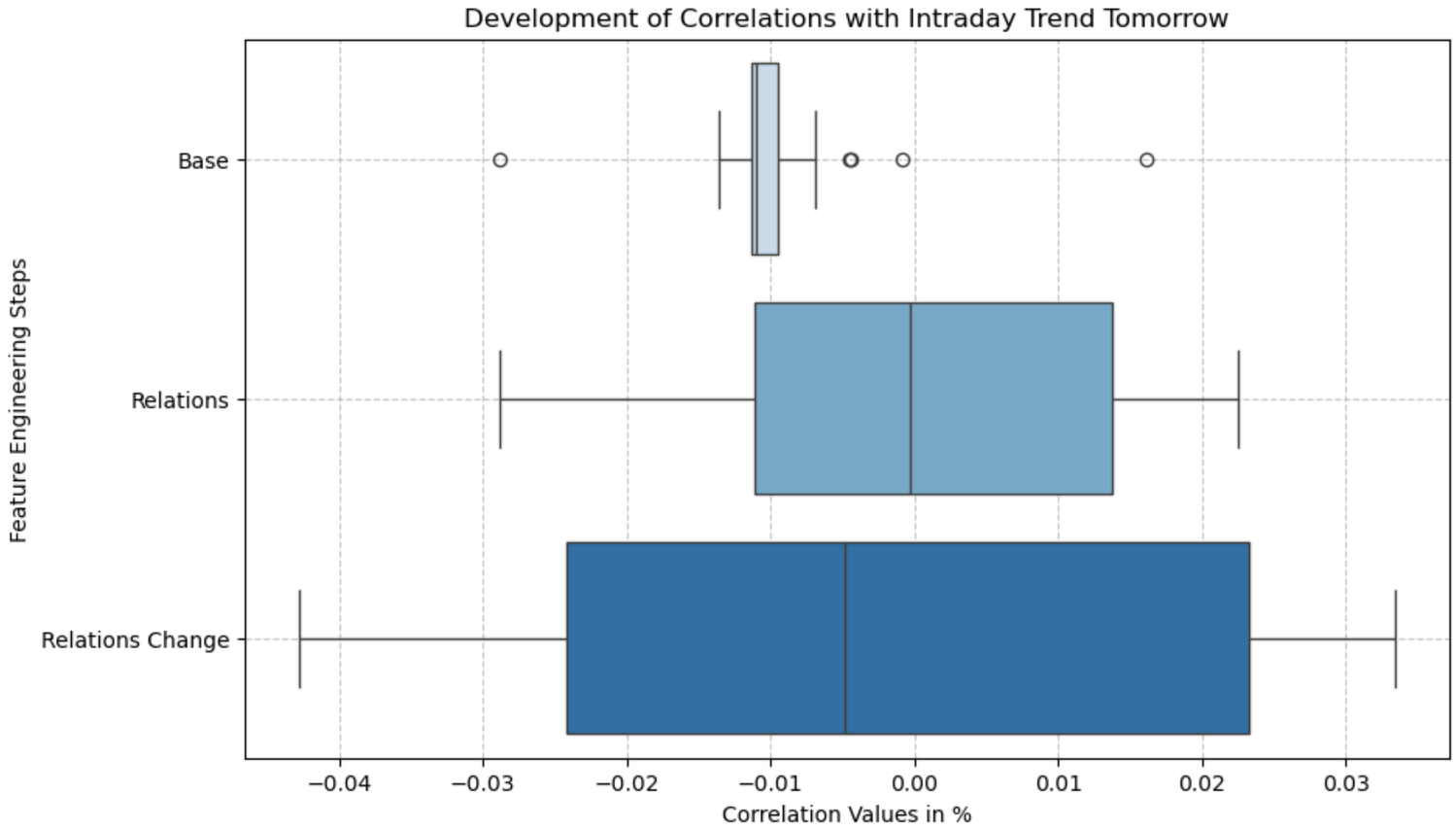
- Close - Open
- Negative —> 0 // Downtrend
- Positive —> 1 // Uptrend



Cleaning and Feature Engineering

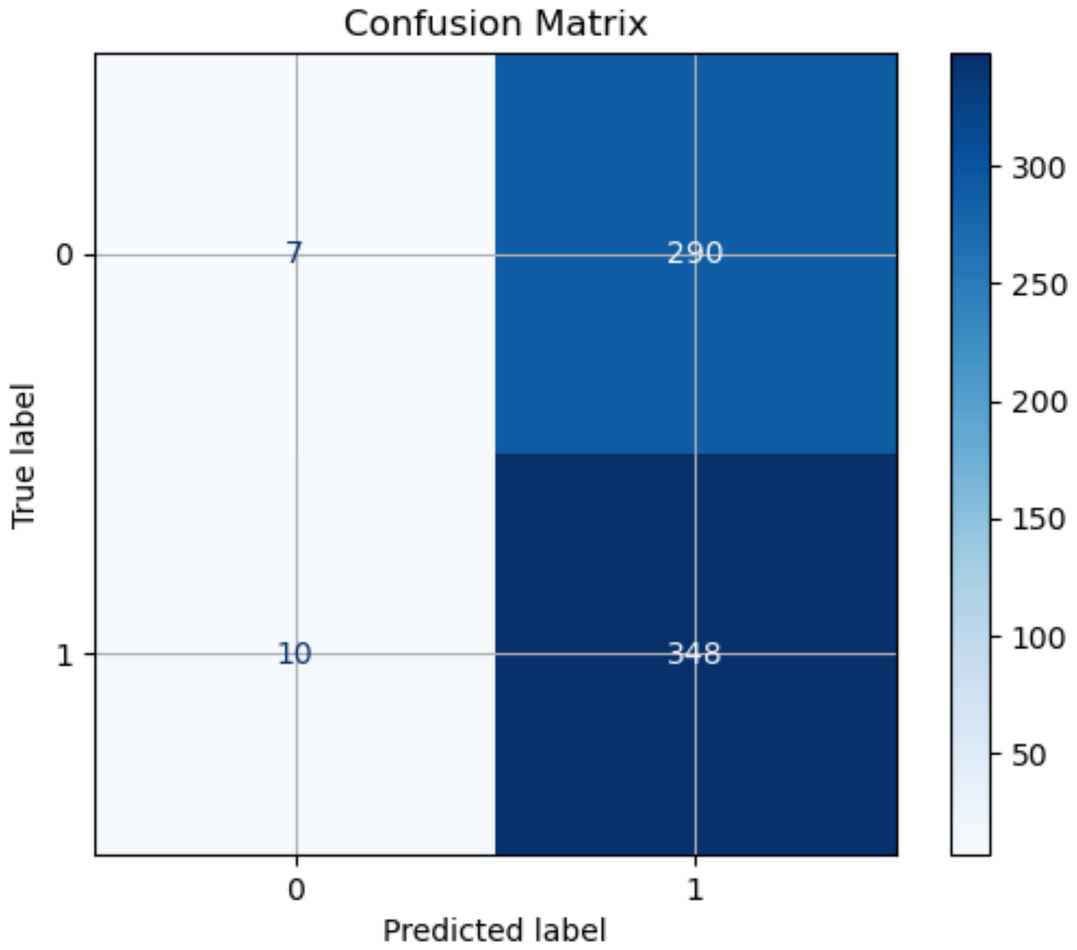
Dropping, merging and creating features

- 3 steps to create better correlations to the target
- Even though the dependencies became more visible, the correlations were still very weak

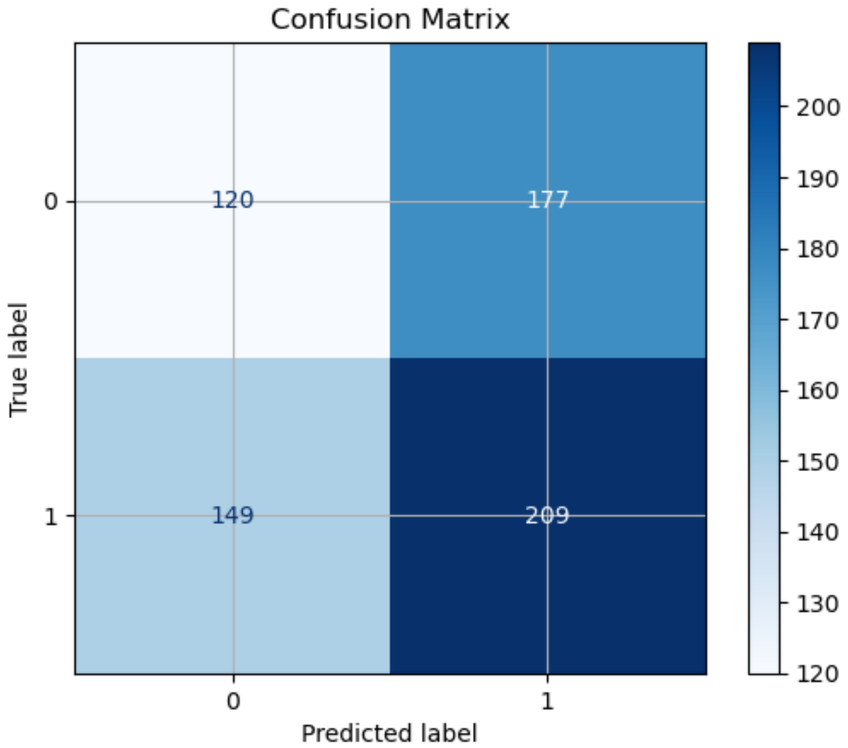


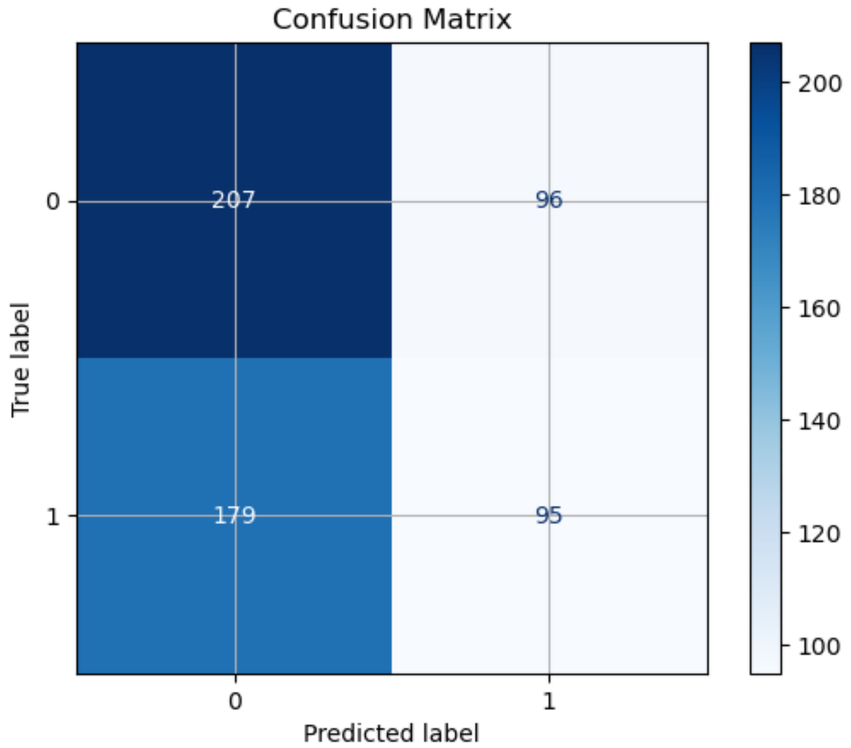
Logistic Regression and Sampling Manipulation

- Model always predicts intraday uptrends
- Different train-test-split, solver, more iterations, tuned features and scaling did not improve the model
- Other models in the LazyRegressor had the same issue
- Assumption: The imbalance in the target is the problem



Logistic Regression and Sampling Manipulation

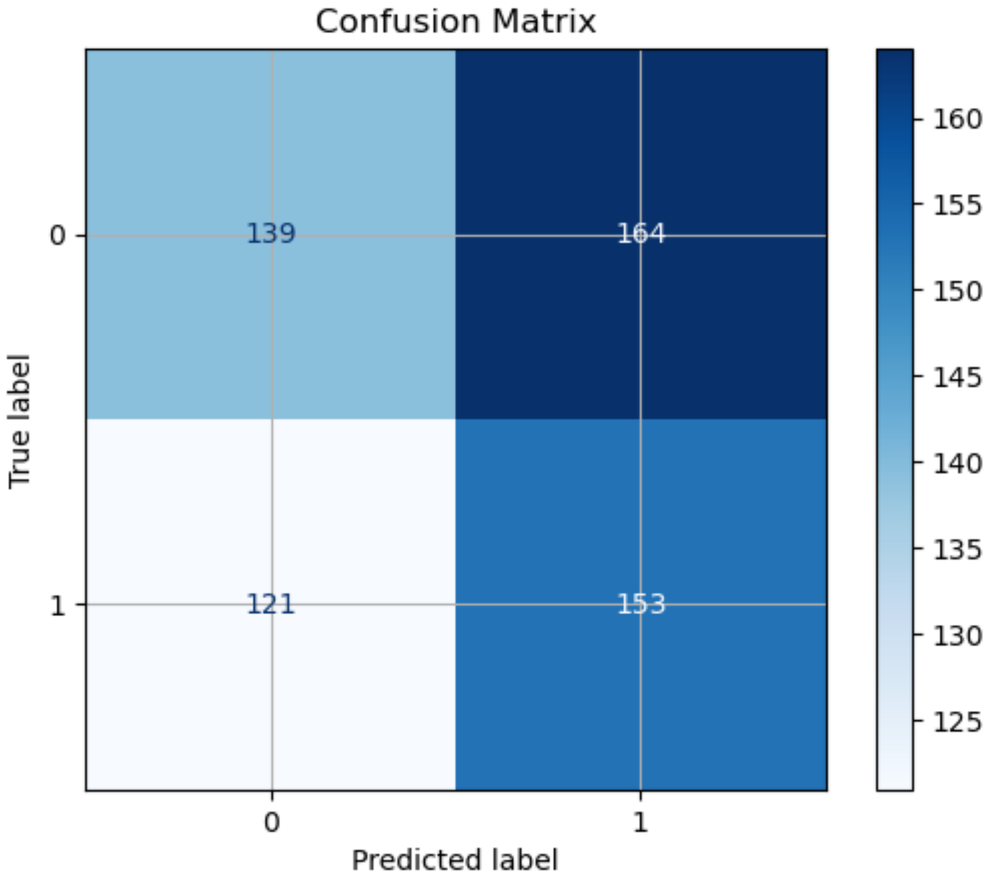




- Oversampling with SMOTE achieved a better distribution, but the results still had too many errors
- Undersampling with TomekLinks achieved good results for the downtrend, but had an overall negative bias

Logistic Regression and Sampling Manipulation

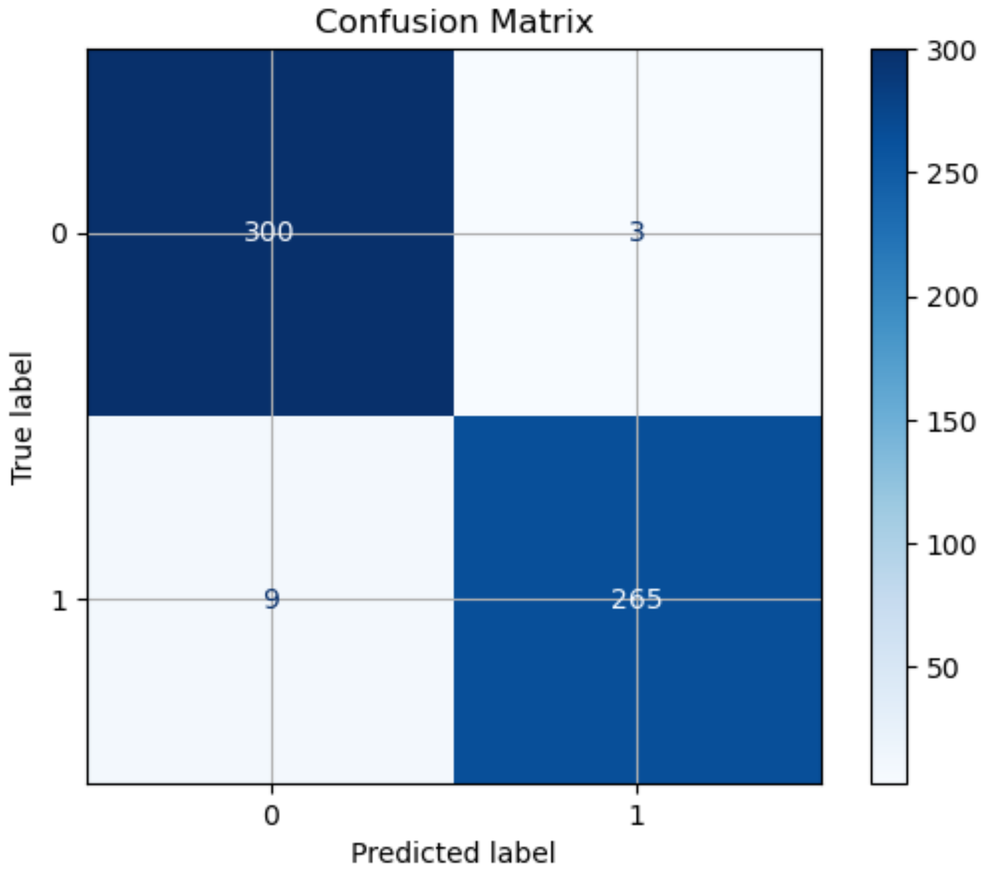
- Introduction of the Voting Classifier to combine two different models / samplings
- Sample with positive (uptrend) bias and sample with negative (downtrend) bias in one model
- Result: Slightly better distribution of errors, but the overall accuracy did not improve



Random Forest Classifier

- After the resampling the Random Forest Classifier achieved great results with a precision of almost 98% and an R2 of 0.66

- A feature importance test showed that the momentum indicator was by far the most important feature for the model (0.8)



Summary

1. The technical indicators correlation to the target value could only be slightly improved

2. Resampling can have a very big impact on a model performance

3. The voting classifier is an interesting tool to combine models

4. While some models like the logistic regression are not useful to tackle this prediction problem, the models in the decision tree family show overall strength for this use case

5. Disclaimer: This is no financial advice. The analysis might have mistakes. For trading you still need a trading and money management strategy no matter how promising an analysis seems to be.

Thank you!

Q&A