# Introduction to Machine Learning

Group 2: Final Project Presentation

---

Jan Heinrich Schlegel, Robert Bibaj, Simon Klaassen, Thomas Meier

April 26, 2022


University of Zurich[UZH]

# Data Cleaning

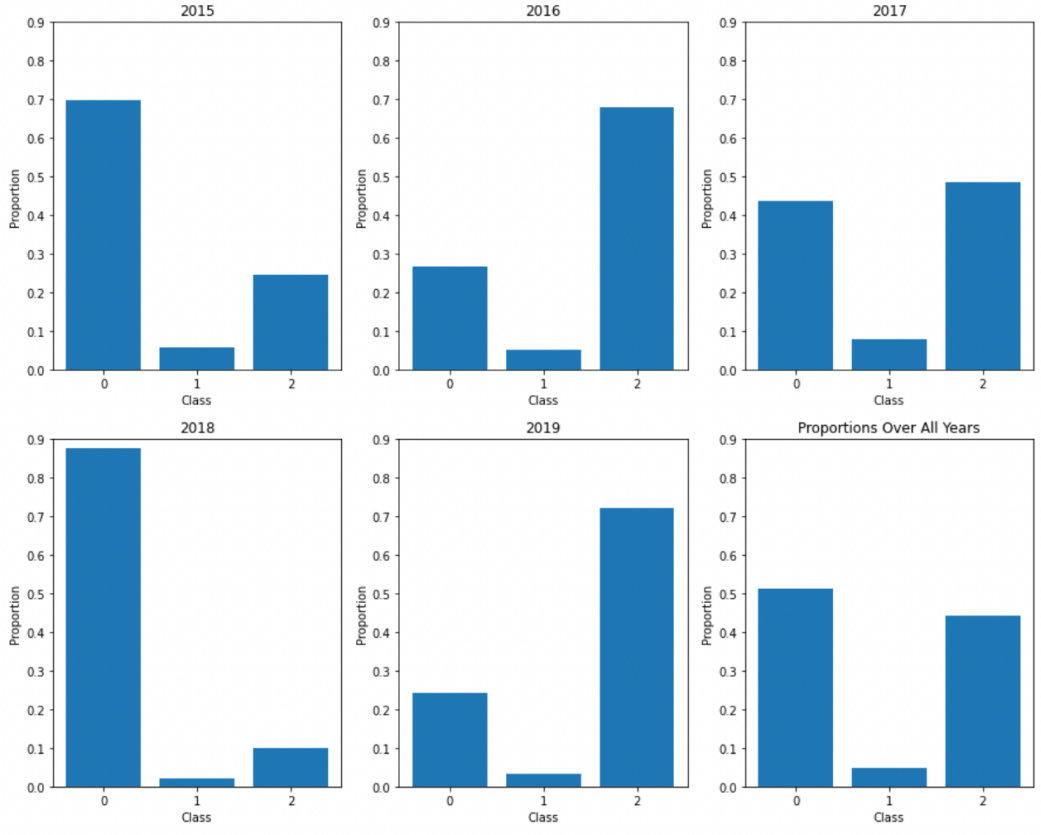


- Concatenating the Datasets: Creating an excess return feature and modifying the existing class feature
- Dealing with NaN's:Threshold of 30%, KNN-imputer
- Dealing with Outliers: Isolation Forest Algorithm, works well with high volume data and can detect data anomalies considering multiple variables

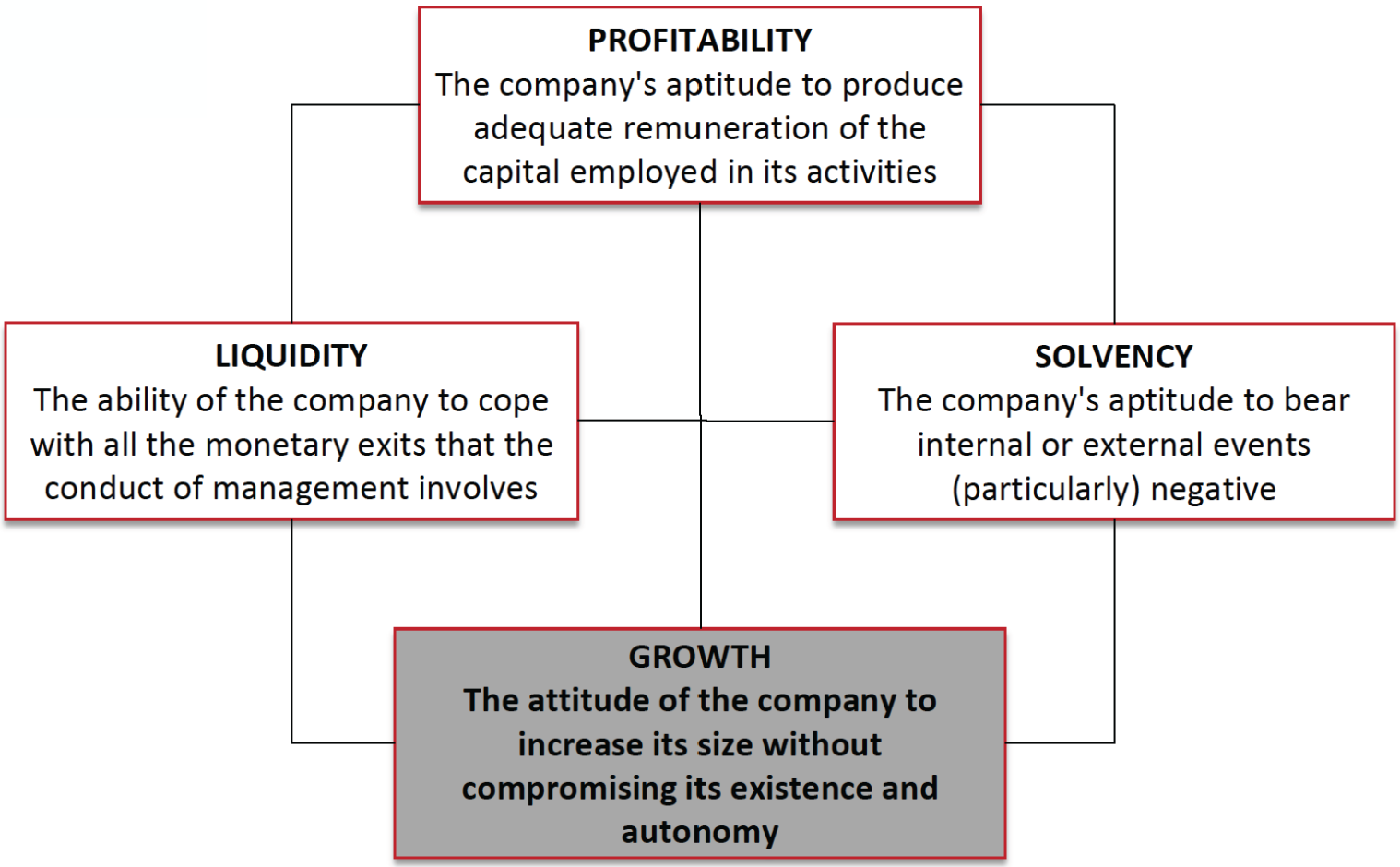
# Class Imbalance



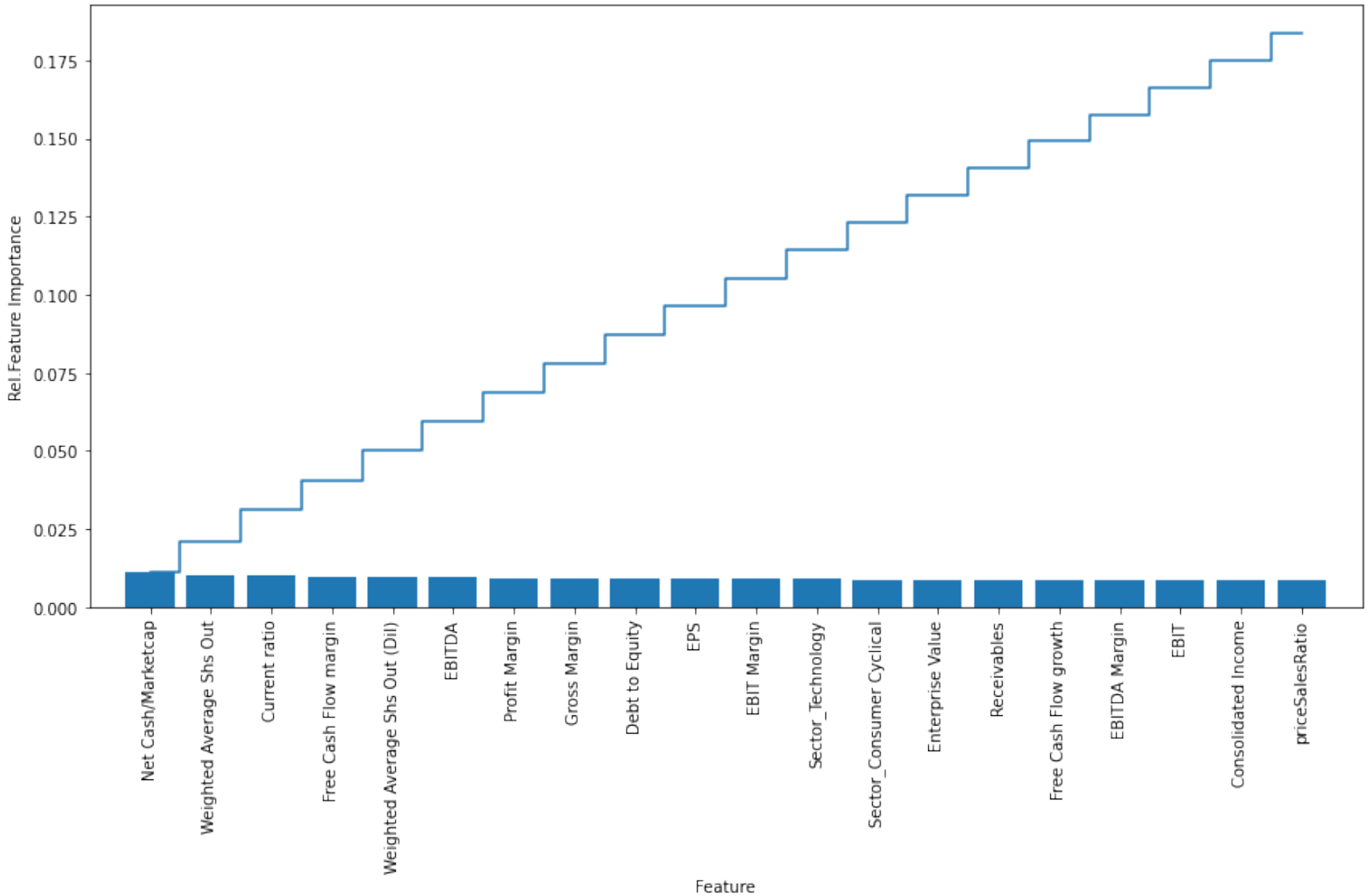
- ▷ SMOTE algorithm and SMOTENC algorithm ⇒ many problems
- ▷ RandomOverSampler as an alternative



## Sustainable Growth Model as Fundamental Analysis Framework



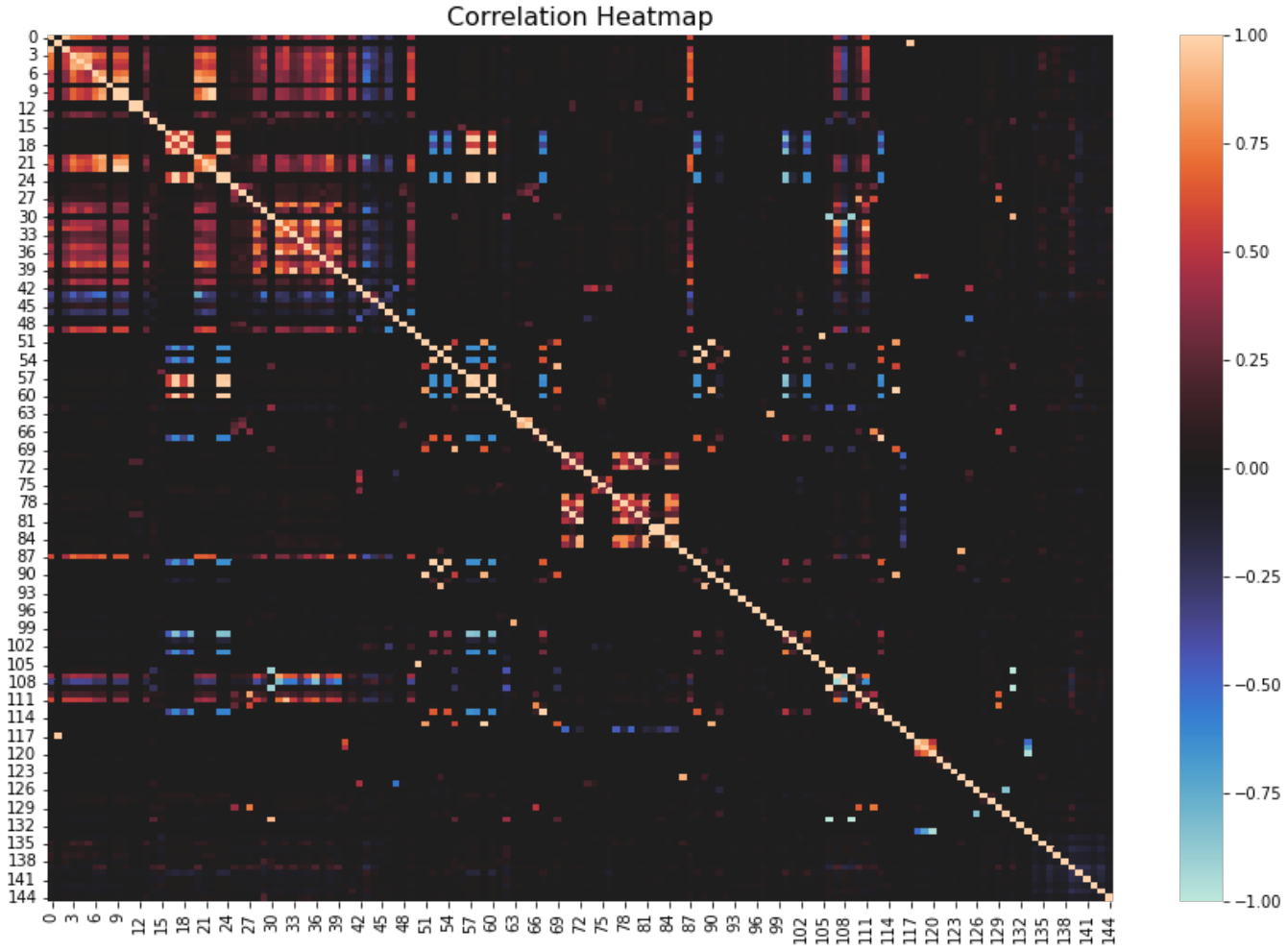
# Feature Selection



# Dimension Reduction

University of Zurich[UZH]

- ▷ Many features are only barely correlated ⇒ PCA problematic
- ▷ Kernel PCA as an alternative



# Results

University of Zurich[UZH]

**Weighted-Averaged $F_1$-Score:**

$$F_1\text{-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The weighted average is calculated by taking the mean of all per-class $F_1$-Scores (calculated in a One-vs-Rest approach) whilst considering the number of occurences of each class.

- ▷ Baseline: Choose class randomly according to class proportions in train set
- ▷ Baseline weighted $F_1$-Score: 0.466

| Approach | Log Reg | Naive Bayes | Random Forest | XGBoost | SVM | Neural Network |
|---|---|---|---|---|---|---|
| All Features | 0.508 | 0.358 | 0.544 | 0.578 | 0.470 | 0.540 |
| Feature Sel RF | 0.440 | 0.390 | 0.492 | 0.470 | 0.416 | 0.535 |
| Feature Sel XGB | 0.466 | 0.505 | 0.494 | 0.446 | 0.273 | 0.536 |
| (Kernel) PCA | 0.456 | 0.508 | 0.507 | 0.526 | 0.459 | - |
| Feature Engineering | 0.513 | 0.463 | 0.535 | 0.581 | 0.493 | 0.531 |

- Do our results imply that the "Efficient Market Hypothesis" does not hold?
- Not necessarily, since we do not know whether the heightened prediction reliability of our algorithms translates into a superior market performance
- Risk of an Omitted Variable Bias: Are our algorithms transferable to other situations e.g. markets in other regions?

# References



Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, *22*(4), 341–352. https://doi.org/10.1080/07388940500339183

Fama, E. F. (1960). Efficient market hypothesis. *Diss. PhD Thesis, Ph. D. dissertation.*

Higgins, R. C. (1977). How much growth can a firm afford? *Financial Management*, *6*(3), 7–16. http://www.jstor.org/stable/3665251

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Lo, A. W., & MacKinlay, A. C. (2011). A non-random walk down wall street. *A non-random walk down wall street*. Princeton University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Schölkopf, B., Smola, A., & Müller, K.-R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, *10*(5), 1299–1319. https://doi.org/10.1162/089976698300017467

Shrive, F. M., Stuart, H., Quan, H., & Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: A comparison of imputation methods. *BMC medical research methodology*, *6*(1), 1–10.

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90. https://doi.org/https://doi.org/10.1016/j.inffus.2021.11.011