Massachusetts Institute of Technology

15.095 Project Final Report

Venture Capital and Private Equity Under a Modern Optimization Lens

AUTHORS

Edoardo Italia
Simon Weill

Supervisor

Vassilis Digalakis

December 9, 2020

Contents

1 Introduction **4**
1.1 Problem Statement . . . 4
1.2 Objective . . . 4
1.3 Data . . . 4
1.4 Exploratory Analysis . . . 5
1.5 Approach . . . 5

2 Models **6**
2.1 Machine Learning Prediction . . . 6
2.1.1 Robust Regression . . . 7
2.1.2 Regression Trees . . . 7
2.1.3 Random Forest . . . 8
2.1.4 XGBoost . . . 8
2.2 Clustering . . . 9
2.3 Optimization . . . 10

3 Results **12**
3.1 Experimental Design . . . 12
3.2 Parameter Sensitivity . . . 12
3.3 Takeaways . . . 13

4 Future Work **13**
4.1 Assumption Improvements . . . 13
4.2 Modelling Improvements . . . 14
4.3 From Model to Product . . . 14

5 Conclusion **14**

References **15**

Appendix **16**

A List of Features **16**

B Varying optimization for different number of clusters **17**
B.1 Optimal Regression Tree . . . 17
B.2 Random Forest . . . 17
B.3 XGBoost . . . 18

C Feature importance for different models **19**
C.1 $k = 5$ clusters . . . 19
C.2 $k = 50$ clusters . . . 19

D Simulation for different investor profiles **20**

Abstract

The venture capital and private equity (VCPE) industry represents a fast-growing sector of investment asset classes. Despite the growing size of this field, the advent of machine learning, and the importance of optimal decision-making in the investment process, the VCPE industry appears to have been dormant to the use of new technologies. Indeed, the vast majority of an investment decision relies on the use of older technologies, and qualitative assessment carried out by experts with inherent biases. In this report we provide an approach leveraging machine learning and optimization to predict investment growth potentials, and optimally invest capital accordingly. Our results confirm the effectiveness of the intersection of these two fields – with our investment decisions returning 1.5 to 5 times more than our baseline model. With a promising warm start, we believe there is a significant opportunity in this sector, with further research required to substantiate our findings.

1 Introduction

1.1 Problem Statement

Poor investment decisions in private markets can lead to serious consequences. Bankruptcies,environmental damages [1], and severe job losses [2] represent a mere fraction of the resulting damage caused by poorly executed investments. Despite the pivotal importance, investment professionals are faced with increasing competition to drive deal execution at faster paces. And in spite of this need, innovation in private market investments has stagnated. This diminished innovation, combined with today's levels of data thus offers a unique opportunity to leverage advanced modeling techniques to augment baseline performances. We believe machine learning through an optimization lens can provide investors with an edge to make better, more conscious decisions.

1.2 Objective

Our objective is to develop advanced machine learning models in order to predict the future performance of an investment and prescribe an investment strategy accordingly. We choose a Markowitz [3] portfolio optimization based on return and risk estimates inferred from our machine learning models and which allocate investments in an optimal way given the fund constraints.

1.3 Data

The data was obtained from Zephyr, an online repository of M&A deal activity in the private markets. The dataset consists of 2,000 deal observations described by 48 variables spanning company financial metrics (EBITDA and Sales) across a number of years prior to and following the investment, as well as company target geographical and sector-based descriptions. Unfortunately, more granular descriptions could not be obtained. Generally speaking, as a result of the nature of the private investment process, target company information is more concealed. Thus, alternative features for company performance - such as online activity, detailed cost and capital structures, and senior management - are not readily available. Given data access constraints, certain initially postulated databases could not be used to obtain additional company descriptors. While gathering such alternative data would have been beyond the scope of this project, it most certainly remains an aspect deserving of additional research. Indeed, based on expert interviews we carried out during the project, advanced analytics in due diligence processes remains a challenge to be fully tackled.

Data cleaning and wrangling was performed to ensure a standardized format for each column. Due to the secretive nature of private market investments, a non-negligible number of deal variable elements were missing in the data set. These were imputed using Interpretable AI's k-NN optimal imputation methods. Feature engineering was instead performed to allow for a better understanding of the patterns in the data set. Geographic regions, for instance, contained too many categorical levels, and were thus clustered into general parent groups - this reduced the total number of geographies from 50 to 4. Certain features were instead engineered to more adequately express additional performance metrics that could not be directly obtained from Zephyr, but are nonetheless industry standard metrics. The feature engineering led to the following general categories of new variables:

- Company profitability metrics (EBITDA margins, etc.)
- Company growth metrics (EBITDA and Sales growth, valuations etc.)
- Investment performance metrics (return on investment, annualized returns, etc.)

1.4 Exploratory Analysis

The distribution of the deals can be found below in Figure 1(Left). Clearly, larger deals, above the €1B are more unlikely as a result of the greater degrees of complexity surrounding the size – review committees, deeper due diligence, and regulatory approval in certain cases [4]. Indeed, a positive correlation can be expected between deal size and investor involvement. Greater capital commitment by an investment team will require more active commitment to ensure adequate value generation – as opposed to smaller ticket sizes, as these represent a smaller share of a fund's active capital.

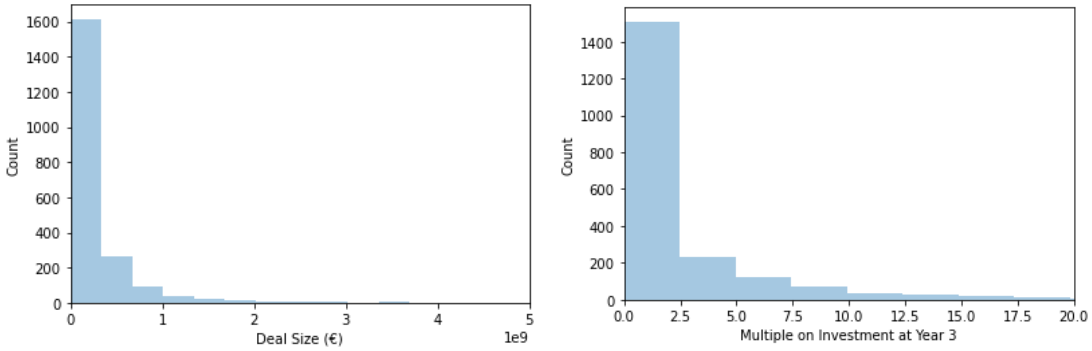

Figure 1: Distributions of investment sizes (in '000s) and multiple for return on investment.

Figure 1(Right), instead, shows the distribution of returns for all deals in the data set. With a considerably high average annualized return of over 40%, these investments significantly outperform the public equity market average of 11%, expected given the higher risk profile of the private equity asset class. Despite the higher returns, there remains potential for additional improvement.

1.5 Approach

The nature of this problem is multi-sided, requiring a number of techniques to be analysed to fully investigate this challenge. We thus tackled the problem in a divide-and-conquer approach by splitting the analysis into two stages (as seen in Figure 2):

1. Investment performance predictions
2. Investment decision prescriptions

.

It is important to stress the novelty - and the resulting uncertainty - of this approach. Current "best practices" merely rely on Excel modeling, with dynamic valuations calculated based on project performance. Indeed, discounted cash flow (DCF) models are

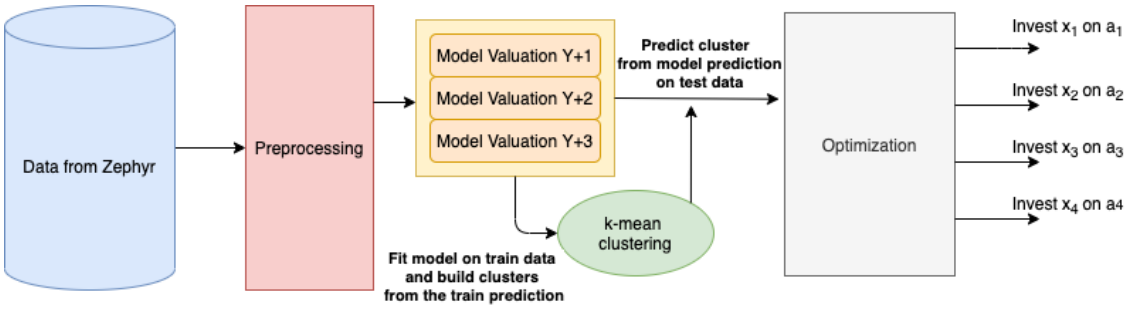

Figure 2: General Architecture of our model. We first predict valuations, then optimize our portfolio based on bucketed predictions and constraints.

typically used to assess the quality of an investment, taking into account a rate of return (ROI) to project valuation performances. While projections require numerical inputs that are obtained after industry expert consultations, an automated, trend-fitting approach is non-existent. Hence, although our approach will attempt to emulate investment decisions conducted by an industry expert, it still lacks key elements that are usually taken into account in private markets, as briefly mentioned in the data section. The details of our modelling approach, both for investment performance predictions and investment decision prescriptions are described in the following sections.

2 Models

2.1 Machine Learning Prediction

To identify the optimal prediction approach, a number of traditional and more advanced approaches were implemented on the engineered training set to predict company valuations across three different years. Specifically, models were trained to predict valuations 1, 2, and 3 years after an investment – each model was trained independently from each other, such that predictions at Year 1 would not be used to predict valuations at Year 2. Robust LASSO regression, Regression Trees (CART), Optimal Regression Trees (ORTs), Random Forests, and Gradient Boosting Methods (XGBoost) were all implemented. In each model instance, grid search procedures were adopted to finetune relevant hyperparameters - a breakdown of lists of the grid values for each model can be found in the Appendix. The reasoning behind the 3-year limit is dual-sided: one the one hand, investments are typically reassessed after 3 years; on the other hand, our database did not contain sufficient information for the following years. The training data for company valuations were calculated by combining each company's EBITDA (Earnings Before Interests Taxes Depreciation and Amortization is a common earning metrics for companies) with its relevant Valuation/EBITDA multiple - a common unit used to benchmark prices paid for investments.

Essentially, our models try to learn the following functions: $F_t : x_i \to y_{t,i}$ where $x_i \in \mathbb{R}^p$ is a given asset i made of p features (full list in Appendix A) and $y_{t,i} \in \mathbb{R}$ is the valuation of the asset i at time t. Given the problem at hand, it is possible to estimate that the valuations will generally increase, especially for fast growing companies. Thus, it is expected that the R^2 values are high. Of more importance, however, is that the correct growth trend is identified, as this will be leveraged in the optimization.

2.1.1 Robust Regression

A linear LASSO regression model was applied as an initial approach to predicting valuation. At its core this is a simple procedure that can lead to powerful results if correctly implemented. Given the inherent uncertainty in the predictions, resulting from the numerous hidden variables that can affect an investment's performance - socio-political, economic, and global healthcare events – robustness is a must. As such, a penalizing term was introduced and tuned to 0.1 to protect against uncertainty. As alluded to above, the obtained R^2 values are consistently high across the three years. Year 3 shows the lowest performance due to the higher associated uncertainty arising from the longer-term prediction.

2.1.2 Regression Trees

CART

Regression Decision Trees were thus implemented as a benchmark for interpretability. CART's top-down approach allows for an increased intuition behind the output predictions, as well as increased performances depending on the data structure and hyperparameters used. Despite its greedy nature, CART is unique in interpretability, and can help categorize more important variables, improving communication with less technically-versed industry members. Implementing CART demonstrates consistently higher performances than LASSO.

Optimal Regression Trees

While Decision Trees can be highly effective and performing, they present a number of limitations. The primary inhibitor is the lack of global optimality that is achieved when training trees; in this regard, optimization methods can be leveraged. To further improve performance without sacrificing interpretability, ORTs were implemented. By modifying the prediction formulation within each tree leaf, ORTs have generally been able to obtain high performances and interpretability. Figure 3 below shows the obtained ORT for Year 3. In this regard, the most important split variables were assigned to the sales growth a year before the investment, as well as 'Sales' and 'ValuationAtInvestment'. These are intuitive results, as these are among the primary primary features assessed when evaluating an investment: its current worth, the profit potential, and the historical growth rates. A collapsed ORT for the Year 1 predictions can be found in Appendix B.1, and shows 'LastEBITDA' as being most important, followed by a sequence of valuation categorizations - likely due to the fact that valuations on a short time frame are more dependent on recent valuations, than longer time frame projections.

From a prediction perspective, ORTs were not found to be particularly high performing. This may be a result of the variability of the data within each leaf. Since each leaf in our ORT outputs a constant prediction based on the mean of the leaf, the number of data points within each leaf will affect the leaf mean. Outliers and variable performances will thus lead to constant values that reduce the out-of-sample R^2 performance - though this is no guarantee of optimization performance.

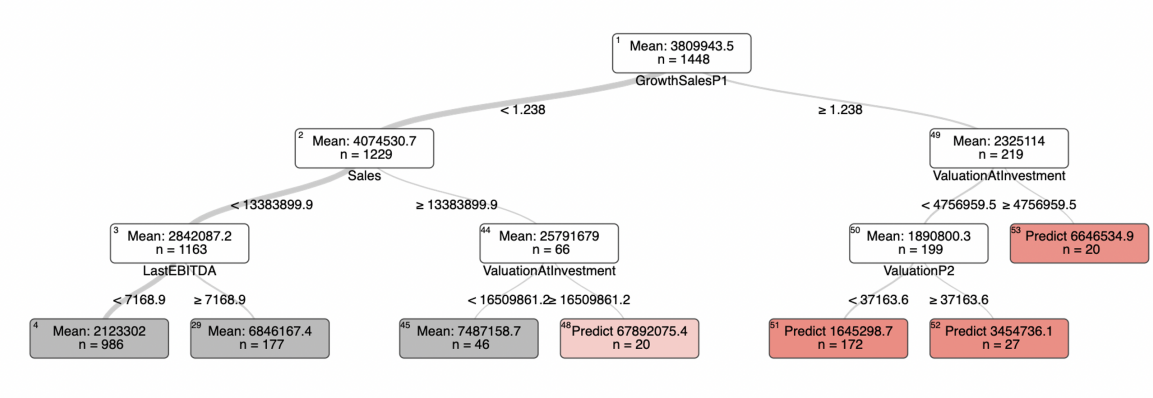

Figure 3: Collapsed Optimal Regression Tree obtained for the valuation prediction Year 3 after investment

2.1.3 Random Forest

Finally, more advanced methods were adopted to enhance performance. While certainly not interpretable, random forest, and gradient boosting approaches have garnered considerable attention as a result of high recorded performances. With an absent decision-making process breakdown, random forest models and gradient boosted models are limited to feature importance plots. The feature importance plot for our Random Forest plot is shown in Figure 4. Notably, the most important variables differ from those observed using ORTs - 'ValuationAtInvestment', 'TicketSize', and 'LastEBITDA' are now the top three. Presumably, the postulated relationship is that of a higher current valuation leading to a higher future valuation. It is possible then that 'TicketSize' and 'LastEBITDA' qualify the investment potential in the target company. A feature importance plot for Year 1 predictions is found in Appendix B.2 - with the same three variables as above classified as most important.

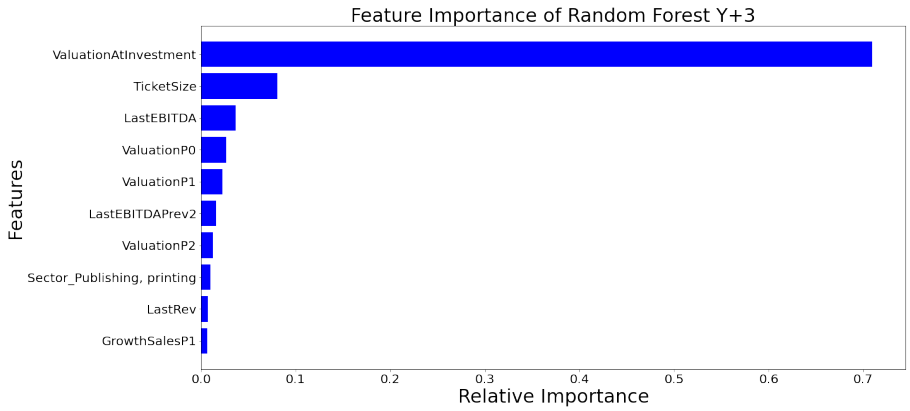

Figure 4: Feature Importance plot for Random Forest model, predicting valuations at Year 3 after investment

2.1.4 XGBoost

To further test alternative approaches, XGBoosting was selected as a result of its known high-performance. Indeed, across all years, XGBoosting provided the highest performances in R^2, with generally lower MAE. It is important to note that while the MAE

may appear excessively high, this is likely due to the presence of certain incredibly high performing investments, which skew the MAE to appear worse than it actually is. Indeed, the majority of the valuation predictions fall within 50% of the true value. A feature importance breakdown for XGB is shown below in the Appendix B.3. The relationships observed are analogous to those found using Random Forest models and reasonably consistent for Year 1 and Year 3 projections. The main difference, however, is in the model's ranking of sales growth parameters; which, echoing the decisions in ORT, is classified as being more important.

Table 1: Summary of the results for each modelling method on every year.

Model	Year +1		Year +2		Year +3	
	MAE(%)	R^2	MAE(%)	R^2	MAE(%)	R^2
Lasso Regression	0.28	0.99	0.32	0.99	1.43	0.74
ORT	0.38	0.55	0.55	0.39	1.71	0.32
CART	0.78	0.95	0.57	0.98	1.72	0.74
RF	1.16	0.99	1.11	0.98	1.61	0.87
XGBoost	0.20	0.99	0.40	0.99	1.54	0.86

2.2 Clustering

After predicting the valuation $\tilde{y}_t,i = F_t(x_i)$, we want to compute the annualized return on investment (ARR) in order to cluster the similarly performing companies together. The main motivation on clustering our predictions based on cluster established on predicted train data is to replicate a leaf-like partition effect from prescription methods for a non-tree based model. Essentially, the method described below provides weights w in the estimation of $\mathbb{E}[\mathbf{y}|\mathbf{X} = \mathbf{x}] \approx \sum_i w(\mathbf{x},\mathbf{x_i})\mathbf{y_i}$ as explained in Chapter 13 of Dunn, Bertsimas [5].

Let's define the ARR of a valuation prediction $\tilde{y}_{t,i}$ as:

$$k_{t,i} = k_t(\tilde{y}_{t,i}) = \left(\frac{\tilde{y}_{t,i}}{y_0}\right)^{\frac{1}{t}} - 1$$

Where y_0 is the valuation at time of the investment so the ratio $\frac{\tilde{y}_{t,i}}{y_0}$ correspond to the return on investment (ROI) at year t, which is annualized with the exponent $\frac{1}{t}$.

We then decide to aggregate our prediction to get the mean ARR, $\mu_i = \mathbb{E}_t(k_t,i)$, and its standard deviation, $\sigma_i = \sqrt{\mathbb{E}_t(k_{t,i}^2) - \mathbb{E}_t(k_{t,i})^2}$. This provides easily interpretable metrics for each of our prediction (risk-return metrics) that can be used in our clustering algorithm.

We then fit k-means clustering to the training predictions and assign test predictions to the different clusters using k-means clustering so companies exhibiting the same type of return behavior or clustered together.

We later use the mean of intra-cluster mean ARR: $\theta_j = \frac{1}{\text{card}(K_j)} \sum_{i \in K_j} \mu_i$ and the mean of intra-cluster standard deviation ARR: $\zeta_j = \frac{1}{\text{card}(K_j)} \sum_{i \in K_j} \sigma_i$ as input for our optimization model such that any investment x_i falling under a cluster K_j, (where $j \in [1,...,m]$, m being the number of clusters) could be described by the pair $(\theta_j,\zeta_j) \forall i \in K_j$.

For simplicity of notation in the optimization part, (θ_i,ζ_i) will correspond to $(\theta_j,\zeta_j)\forall i \in K_j$ such that it reflects the "centroids" of the cluster to which the point x_i belongs.

As per explanations above, the predicted valuations were used to obtain ROIs and ARRs for each year. These were then averaged and clustered into groups of investable companies. Mean ARR and standard deviation with each group were then calculated to obtain the risk-return profiles for each group. This can be found in Figure 5, showing the correlation between the expected return of an investment (cluster mean) and its corresponding associate risk (standard deviation of returns within cluster point).

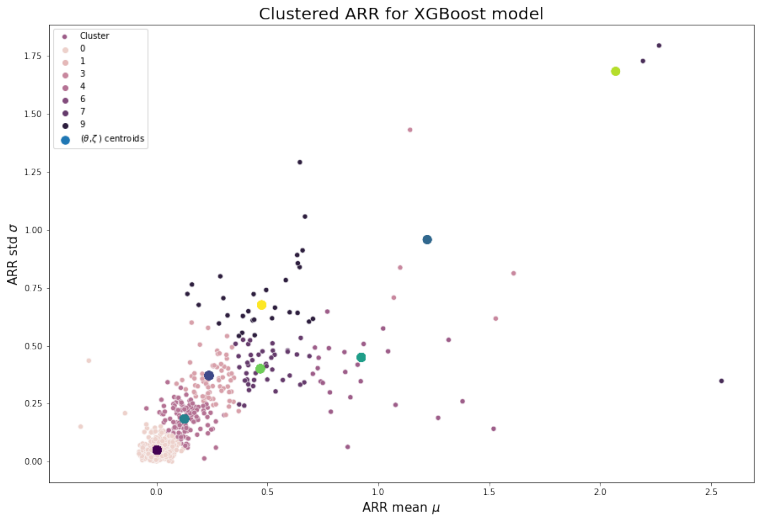

Figure 5: Example of distribution of μ and σ for all companies in the test sets (clusters built based on the training data). The larger points represent the "centroids" (θ_j,ζ_j) of each cluster K_j. Here $m = 10$ but only 7 clusters are visible as no points from the test set have been assigned to the remaining 3.

2.3 Optimization

To model optimal investment decisions, a typical Markowitz [3] portfolio optimization was chosen, where the objective function seeks to maximize the risk-return tradeoff by analyzing mean predicted return, and return standard deviation. This last term can then be prefaced by a penalty term, used to represent the risk appetite each investor may have.

The investment optimization will then further take into account investment fund and investment type constraints specific to each profile. For instance, particular investors may prefer varying degrees of industrial or geographic diversification, resulting from niche investment strategies. Other firms instead, may be faced with budget and investment constraints, as a result of the firm's investment timeline and dynamics. The considerations above lead to the following MIO formulation.

Consider an optimization problem where we try to allocate a budget B into n different companies. Companies are described by their clustered return and risk measure $\boldsymbol{\theta} \in \mathbb{R}^n$ and $\boldsymbol{\zeta} \in \mathbb{R}^n$ and the trade-off between return and risk is controlled by γ (we assume independence between companies' risk).

More over each of the target company appears in a given region and a given sector. To model these two parameters, we use the matrices $R \in \{0,1\}^{n,r}$ and $S \in \{0,1\}^{n,s}$ where r is the number of distinct selected regions, s is the number of distinct selected sectors

and R_i and S_i are the rows of R and S which are equal to 0 for every column except the region/sector of the company i where it equals to 1. Hence by using $\sum_s S_{i,s} = 1$ if the company is within the selected sectors and 0 otherwise, we could select the eligible companies for our portfolio. The same applies for regions. We introduce two sets of decision variables: $\mathbf{x} \in \mathbb{R}^n$, which corresponds to the amount invested in each company and $\mathbf{z} \in \{0,1\}^n$, which corresponds to the binary decision of investing in a company.

We also consider fund constraints, such as the number of investment to make: $a \leq \sum_{i=1}^{n} z_i \leq b$ (we make between a and b investments), a constraint on the maximum amount of money to invest: $\sum_{i=1}^{n} x_i \leq B$ and finally the fact that the ticket size (percentage of the company to acquire) depends on the valuation of the company at time of investment (smaller company would sell a higher percentage of their ventures, whereas bigger ones would only sell a small stake). These minimum and maximum investments could be modelled by two vectors: $U_- \in \mathbb{R}^n$ and $U_+ \in \mathbb{R}^n$, the respective lower and upper bound of money to invest for each company (pre-calculated using heuristic based on percentage stake to sell given the valuation at investment). The overall formulation could be formulated as follow:

$$
\begin{aligned}
\max_{\mathbf{x}} \quad & \mathbf{x}^\top(\boldsymbol{\theta} - \gamma\boldsymbol{\zeta}) \\
\text{s.t.} \quad & a \leq \sum_{i=1}^{n} z_i \left(\sum_r R_{i,r} \sum_s S_{i,s} \right) \leq b \\
& 0 \leq \sum_{i=1}^{n} x_i \leq B \\
& U_{-_i} z_i \left(\sum_r R_{i,r} \sum_s S_{i,s} \right) \leq x_i \leq U_{+_i} z_i \left(\sum_r R_{i,r} \sum_s S_{i,s} \right) \quad \forall i \in [1,...,n]
\end{aligned}
$$

We observe that the parameters of the optimization models are:

- x, the amount invested in each company
- z, 1 if investment selected, 0 otherwise
- B, the overall budget
- S, the matrix of selected sectors
- R, the matric of selected regions
- γ, the risk factor
- a, the minimum number of investment to make
- b, the maximum number of investment to make
- U_- and U_+, who are parametrized by our heuristic (map of which percentage depending on the valuation at investment)
- θ and ζ who are parametrized by:
  - F_t $\forall t$, the learned functions from a given model
  - m, the number of clusters chosen

In the next section we explore the impact parameter variation has on our results.

3 Results

3.1 Experimental Design

The performance of the group of investment decisions was evaluated as a function of the amount of money returned 3 years after the investment is made. To assess the capital allocation performance for each of the five machine learning models a series of inputs were varied in an attempt to emulate the different types of investment firms. The groups of emulated investment firms were categorized as larger and smaller investors, based on budget capabilities (thus the average available investment size), as well as the investor risk profile. As a result, the sensitivity was evaluated as a function of the penalizing term, γ, the total budget available, as well as the minimum and maximum number of executable investments. We also use a baseline model, which makes greedy investments on the companies with highest valuation at investment and an "ideal" model, which invests in the companies with highest ROI on the third year.

3.2 Parameter Sensitivity

Figures 6 (below), 10 (Appendix C.1.), and 12 (Appendix C.2.) show the breakdown of model performances across the variables outlined above. It is clear that machine learning under optimization can be leveraged to deliver superior returns. With both more advanced and more interpretable models significantly improving on baseline models, with our proposed model appearing to return between 1.5 to 5 times higher returns. Notably, once the minimum number of investment surpassed 2, the performance remains virtually unchanged, proving robustness across different investment activity levels. Further, varying the risk factor, γ, incurs variations in performance, though no relationship can be inferred - this may be due to the concept of volatility being more intrinsic to stocks than private equity companies, suggesting that alternative risk metrics should be assessed. The number of clusters used in the pseudo-prescription leaves was also altered. Increasing the number of clusters, reduces the number of points in each leaf, which impacts the variability in mean predicted values.

Varying the investment budget, instead, leads to a curious result. That is, the general convergence of all models towards a constant similar rate of return. This last trend correlates with those observed in the industry, whereby higher active fund amounts lead to diminishing returns.

To benchmark the performance of our approach, we simulated groups of investors, varying sizes and risk appetites. These were categorised into "Small", "Medium", and "Big" archetypes. The annualised returns of these archetypes were calculated by subsetting the dataset to match the archetype constraints, and calculating the annualised returns based on the given ROI at Year 3. The results are shown in Appendix D. While the Small Investor and Medium Investor archetypes were beaten by our approaches, the Big Investor archetype posted higher returns. Larger investments are generally harder to execute and require further work to ensure that the committed capital adequately generates value [6], [7]. In other words, the superior performance of larger investments cannot be solely attributed to the profitability of the target business and its growth prospects. It may be a result of additional decisions carried out by the investment committees and target company management. As such, the higher rates of return for the Big Investor category are expected to beat our models - and this was the case. Note: the better-than-ideal performance is observed simply due to the way in which the Big Investor returns

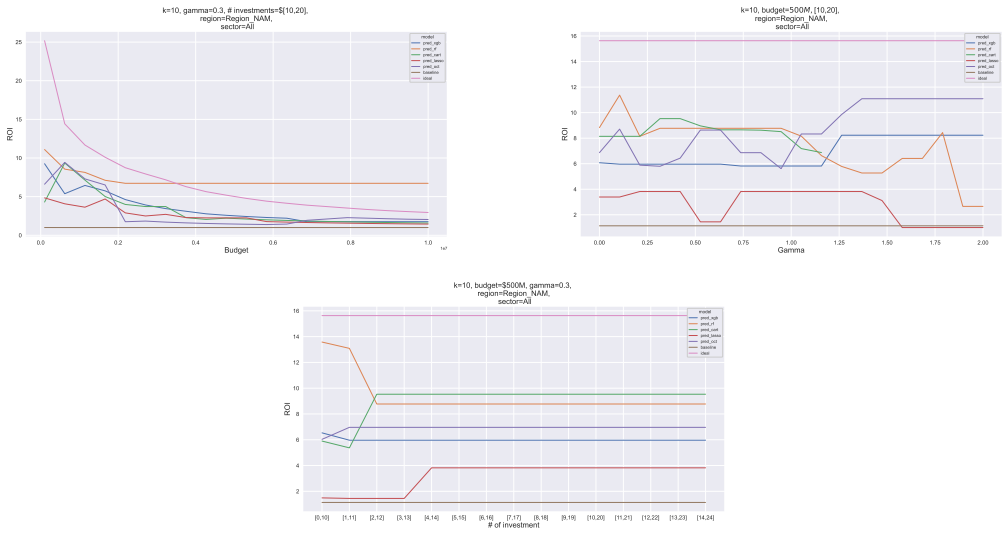

Figure 6: Return on investment for different budgets B, different risks γ and different number of investments $[a-b]$ for $k = 10$ clusters. Note that when the "ideal" is outperform, it is because the other method does not allocate the full budget hence keeps a higher ROI.

were calculated - unconstrained, thus able to allocate more capital than the oracle in the investment.

For machine learning and optimization models to capture a truly holistic investment process would require a significantly enhanced data set with information on target management, executive decisions made, and the outcomes. Unfortunately, such exercise would go far beyond the scope of this project– though should most certainly be evaluated in the future.

3.3 Takeaways

While it is not possible to claim constantly higher returns with the model above, since deeper analyses are required, this proposed model can be used to aid private equity investors in their investment decisions. Nonetheless, the results imply that there may be substantial value left on the investment table. Financial repercussions aside, greater value generation would allow for companies to grow faster, leading to higher employment and greater investments in research and development for society as a whole.

4 Future Work

4.1 Assumption Improvements

The results provide a promising outcome; however, it is important to evaluate the approach used objectively. In conversations with industry experts, the complexity behind a company's valuation was highlighted. Numerous methodologies are often used in a valuation process, ranging from cash flow analysis, to precedent transactions, and comparable company analysis. This last element is where profitability and enterprise value multiples play a role. As such, while our proposed models make use of some valuation methodologies, they fail to capture the full scope of the valuation process.

Future models may, as such, take into account alternative data sources to complement company financials with additional business metrics (customer lifetime value, customer acquisition costs, etc.) as well as macroeconomic factors (industry growth, geo-political trends). Similarly, more advanced investment allocation methods, such as Optimal Prescription Trees, could be evaluated.

4.2 Modelling Improvements

The investment allocation approach must also be considered. The optimization was carried across a testing set of a size (n = 500) that does not fully represent the number of companies that larger funds are capable of evaluating simultaneously. Indeed, the proposed MIO above would need to be revisited based on specific fund aspects, in order to create a more tailored approach to the optimization – this may well present opportunities for future work.

Additional future work, may consider the evaluation of various risk metrics. Covariances, for instance, as well as cluster-wide standard deviation (instead of cluster mean of standard deviations) may be used to provide a more holistic consideration of risk when modeling the investment decision.

4.3 From Model to Product

Our innovative approach grants the need to apply a more entrepreneurial perspective on the problem at hand, thus exploring ways to apply the methods above. Indeed, our approach could be structured as an advanced analysis tool for investment directors - with our product leveraging internal and public data sets to better identify investments. As a result of this potential, we plan on engaging further with industry experts to fully explore this opportunity.

Finally, packaging our code and creating an intuitive user interface would turn our prototype into an actual product for non-technical users. It will enable investors to tune the models to fit their own strategies intuitively to simulate optimally their investment allocations.

5 Conclusion

By taking into account assumptions about company valuations, we have applied advanced machine learning models to predict investment performances in private markets. We then intersected these models with formulations for investment optimizations. Our work has confirmed the possibility to obtain higher levels of value creation, in an industry that has yet to adopt advanced analytics. Alternative data, as well as more advanced models, may well be implemented to augment the quality of the predictions. Furthermore, alternative optimization formulations may be developed to take into account investment deal features that were not included in the analysis above, such as contract terms usually negotiated by the involved parties. We are confident in the industry future use of approaches like these and look forward to the evolution of the investment decision process in venture capital and private equity.

References

[1] Hugh MacArthur, Chris Bierly, Jean-Charles van den Branden, Iwona Steclik, Johanne Dessard and Axel Seemann, *Investing with Impact: Today's ESG Mandate in Private Equity.* Bain Report, 24th of February 2020.

[2] Eileen Appelbaum, Andrew W. Park, *How Private Equity Ruined a Beloved Grocery Chain.* The Atlantic, 16th of February 2020.

[3] Harry Markowitz, *Foundations of Portfolio Theory.* Journal of Finance, 1991.

[4] Sebastian Barling, Omar Salem, *Private equity firms are facing the prospect of bank-style regulation*, https://www.linklaters.com/en-us/insights/publications/2020/march/private-equity-firms-are-facing-the-prospect-of-bank-style-regulation, 2020.

[5] Dimitris Bertsimas and Jack Dunn, *Machine Learning Under a Modern Optimization Lens* Dynamic Ideas, 2019

[6] Christopher Schelling, *The Truth About Private Equity Fund Size* Institutional Investor, 9th of December 2019

[7] Miriam Gottfried, *Private-Equity Firms Are Raising Bigger and Bigger Funds. They Often Don't Deliver.* Wall Street Journal, 18th of June 2019

Appendix

A List of Features

We use the following features for all of our models (PN or Prev$N$ indicates the metrics at year $-N$:

- Region APAC
- Region EU
- Region MEA
- Region NAM
- Sector Chemicals, rubber, plastics, non-metallic products
- Sector Construction
- Sector Education, Health
- Sector Food, beverages, tobacco
- Sector Gas, Water, Electricity
- Sector Hotels & restaurants
- Sector Insurance companies
- Sector Machinery, equipment, furniture, recycling
- Sector Metals & metal products
- Sector Other services
- Sector Post and telecommunications
- Sector Primary Sector (agriculture, mining, etc.)
- Sector Public administration and defence
- Sector Publishing, printing
- Sector Textiles, wearing apparel, leather
- Sector Transport
- Sector Wholesale & retail trade
- Sector Wood, cork, paper
- DealType IPO
- DealType Minority stake
- CompletedDate
- Stake
- PreRevMult
- PreEBITDAMult
- Sales
- EBITDA
- LastRev
- LastRevPrev1
- LastRevPrev2
- LastEBITDA
- LastEBITDAPrev1
- LastEBITDAPrev2
- EBITDAMult
- SalesMult
- ValuationP0
- ValuationP1
- ValuationP2
- ValuationAtInvestment
- GrowthEBITDA
- GrowthEBITDAP1
- GrowthSales
- GrowthSalesP1
- MarginEBITDAPrev1
- MarginEBITDAPrev2
- MarginEBITDA
- TicketSize

B Varying optimization for different number of clusters

B.1 Optimal Regression Tree

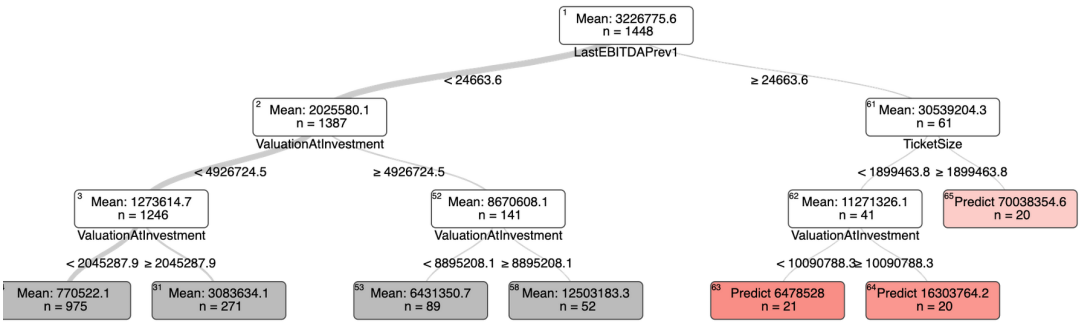

Figure 7: Collapsed ORT for Year +1

B.2 Random Forest

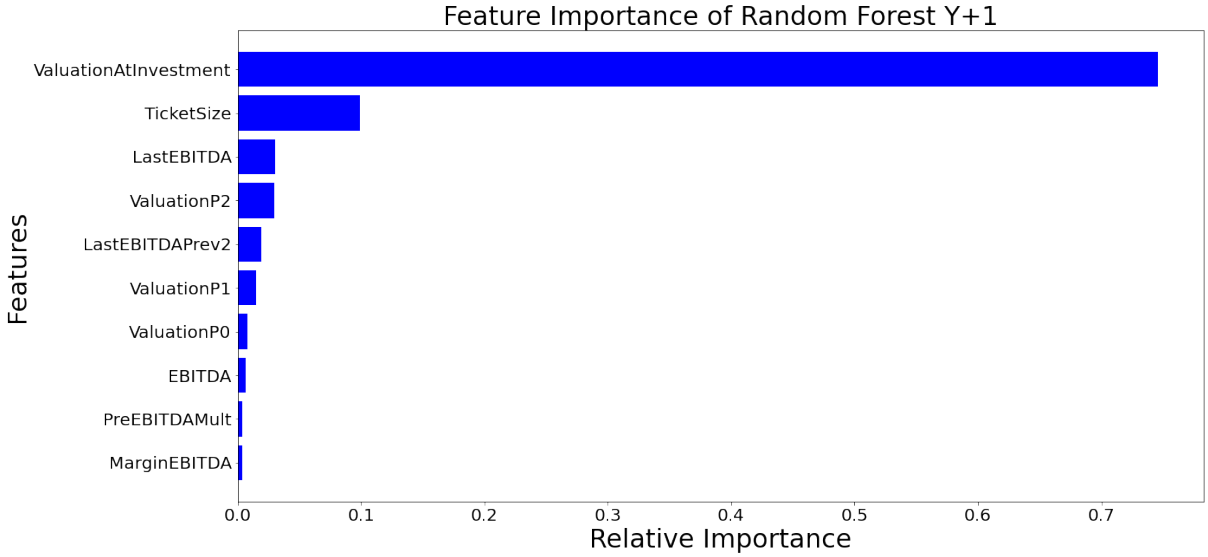

Figure 8: Top 10 feature importances for Random Forest Year 1 after investment

B.3 XGBoost

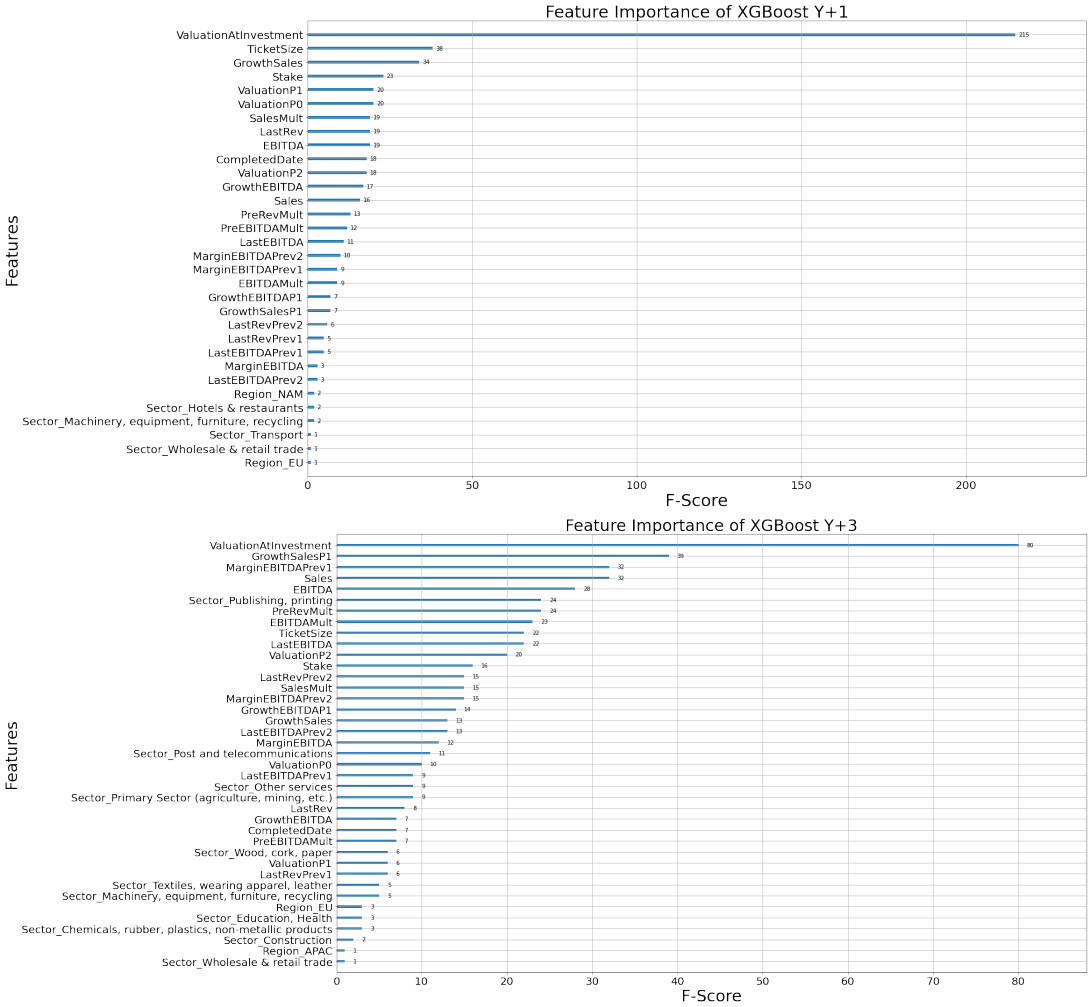

Figure 9: Feature importances for Random Forest Year 1 and 3 after investment

C Feature importance for different models

C.1 $k = 5$ clusters

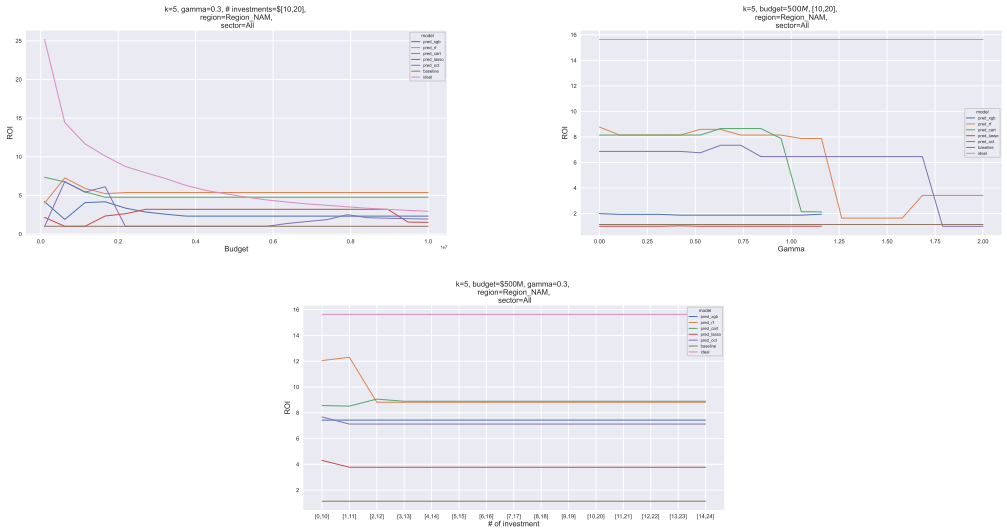

Figure 10: Return on investment for different budgets B, different risks *gamma* and different number of investments $[a - b]$ for $k = 5$ clusters. Note that when the "ideal" is outperform, it is because the other method does not allocate the full budget hence keeps a higher ROI.

C.2 $k = 50$ clusters

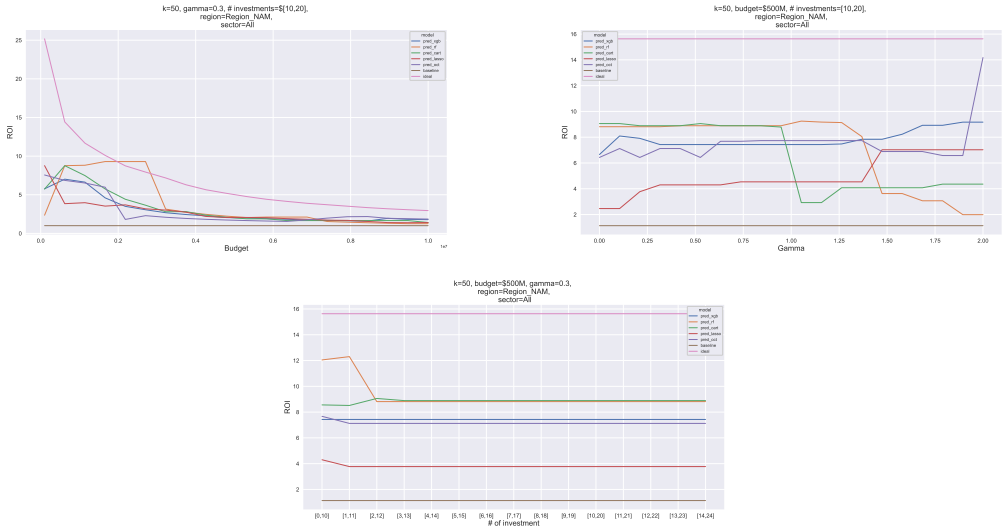

Figure 11: Return on investment for different budgets B, different risks *gamma* and different number of investments $[a - b]$ for $k = 50$ clusters. Note that when the "ideal" is outperform, it is because the other method does not allocate the full budget hence keeps a higher ROI.

D Simulation for different investor profiles

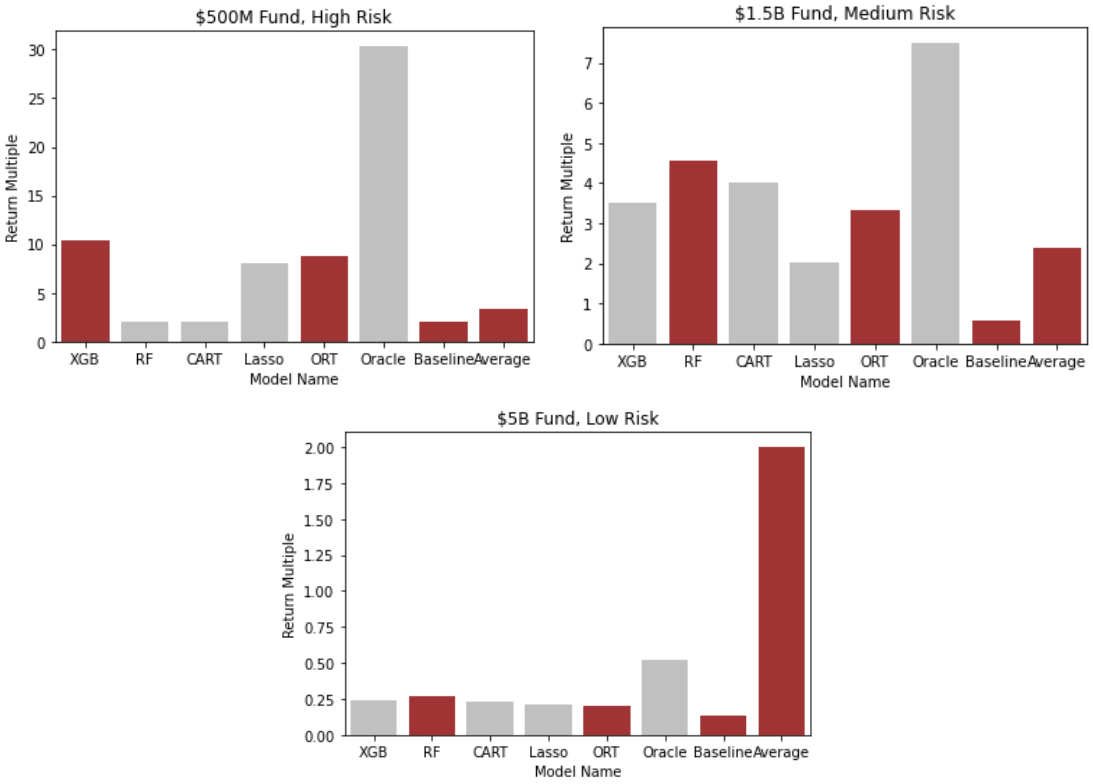

Figure 12: Simulation of returns for different investor profiles. Small, Medium, and Big, correspond to $500M, $1.5B, and $5B in yearly commitments. 'Average' returns taken as average ROI at Year 3 for the specific investor archetype. Note that when the "Average" outperforms all, it is because it does not take into account the investment constraints the oracle is subjected too - there is no limit to the return.