Global Investment Destination Prediction: Leveraging LSTM and Logistic Regression for Optimal Country Investment Recommendations

Lukas Burtscher Master's Program in Data Science
Supervised by: Dr. Michael Wolfesberger
Co-Supervised by: Univ.Prof. Dr. Allan Hanbury

Abstract: This paper explores the predictive modeling of investment opportunities for businesses seeking international expansion. Leveraging extensive datasets sourced from Orbis Crossborder, FDIMarkets, and Refinitiv, our study analyzes investment patterns and country connections worldwide. We employed state-of-the-art methodologies, including Long Short-Term Memory (LSTM) and logistic regression models, developed using the Keras TensorFlow and sklearn libraries, respectively. Hyperparameter tuning and regularization techniques were employed to optimize model performance. Our analysis yields actionable recommendations for investment opportunities, considering historical trends and company-specific factors. The LSTM model demonstrates a nuanced understanding of individual company dynamics, while the logistic regression model emphasizes historical country connections. Through this study, we contribute to the advancement of predictive modeling in financial decision-making, offering valuable insights for stakeholders navigating global investment landscapes.

Keywords: Foreign Investment, Country Selection, Predictive Modeling, LSTM Network, Logistic Regression

1 Introduction

Foreign Direct Investment (FDI) has played a significant role in the global economy since the post-World War II period [1]. Initially concentrated among developed economies, the trend shifted in the 1980s when many Latin American countries began liberalizing their economies to attract FDI, following severe debt crises [1]. This trend continued with China's economic opening in 1985, subsequently followed by other Asian countries, including India's adoption of liberalization, privatization, and globalization (LPG) in the early 1990s [1]. These reforms aimed to attract FDI to address currency crises, stimulate employment, and drive overall economic growth. Developing economies have continuously adapted their economic policies to attract more FDI, recognizing its role in acquiring advanced knowhow and knowledge.

Artificial Neural Networks (ANNs), particularly Long Short-Term Memory (LSTM) networks, have emerged as powerful tools for analyzing complex data [2]. LSTM, a specialized form of Recurrent Neural Network (RNN) [3], excels in learning long-term dependencies, addressing issues like the vanishing gradient problem, and effectively modeling dynamic sequences [4]. With LSTM's ability to retain memory over extended periods, it is well-suited for forecasting and addressing challenging sequence problems, offering state-of-the-art results in predictive modeling.

This paper presents an innovative approach utilizing both Long Short-Term Memory (LSTM) and Logistic Regression models to predict optimal foreign investment destinations for companies. We delve into the dataset employed, detailing the preprocessing steps and model construction process to derive meaningful predictions. The Project can be found on GitHub[1].

In tandem with LSTM, Logistic Regression serves as a benchmark, offering a baseline comparison against the more advanced LSTM model. Despite its simplicity, Logistic Regression provides valuable insights into the predictive performance of the LSTM model, aiding in the assessment of its efficacy in foreign investment prediction tasks.

The output of both models yields a score ranging from 0 to 1, representing the likelihood of a country being an optimal investment destination. Leveraging these scores, we rank the top five foreign countries, providing actionable insights for companies seeking to allocate their resources effectively in the global market.

2 Datasets

The dataset utilized in this study is sourced from Orbis Crossborder, FDIMarkets, and Refinitiv, amalgamating a comprehensive repository of foreign subsidiaries worldwide [5]. Dr. Wolfesberger previously curated and transformed the dataset [5], facilitating our analysis. Two distinct datasets were employed for our investigation.

The first dataset encompasses information regarding investment opportunities analyzed by companies, detailing crucial attributes such as operating revenue, industry classification, total assets, among others. This dataset features the source country of the company and the potential destination country for investment consideration. Additionally, a label column indicates whether the investment was executed or not.

The second dataset focuses on various country connections, presenting details on source and destination countries alongside additional information such as trade agreements, conflicts, cultural distance, and other pertinent measurements derived from established indices.

2.1 Preprocessing

Data preprocessing is a crucial step in machine learning as it transforms raw data into an understandable format. It's particularly important for numerical data and models like LSTM networks and logistic regression for several reasons:

- **Handling Missing Values:** Real-world data often has

[1] https://github.com/lukasthekid/investment-locations-lstm Last Access 19.03.2024

missing values. Data preprocessing can fill these gaps through various methods, such as using the mean, median, or mode [6].

- **Feature Scaling:** Many machine learning algorithms perform better when numerical input variables are scaled to a standard range. This includes algorithms that use a distance measure, like k-nearest neighbors (KNN), and linear regression and neural networks that use gradient descent [6].
- **Encoding Categorical Variables:** Machine learning models require inputs to be numerical. Preprocessing includes converting categorical variables to numerical form [6].
- **Improving Efficiency:** Preprocessed data can speed up the training process by reducing the complexity and computation time of machine learning algorithms [6].

For numerical features, we decided to remove the mean and scale to unit variance. This is done by applying the following formular to the features:

$z = \frac{x-u}{s}$

where: x is the feature to be scaled, u is the mean of the training samples, and s is the standard deviation of the training samples. In the end, we joined the two data frames based on source and destination country. The final dataset then includes company and country information. We got 366413 observations between 2014 and 2019 with 731 features.

3 Methods

The label column in our dataset indicated whether an investment opportunity was seized (1) or not (0). Given the imbalance in our dataset, with only 3267 observations labelled as 1, we decided to optimize our model based on the Area Under the Curve (AUC) score instead of binary accuracy. AUC considers all possible classification thresholds, unlike accuracy which considers a single threshold. This makes AUC a better metric when there is an imbalance in the classes [7]. AUC is less sensitive to datasets with an imbalance between the positive and negative classes12. For example, in a dataset with 99% negative instances, a model that always predicts "negative" would achieve 99% accuracy. However, such a model would have an AUC of 0.5, which is no better than random guessing12 [7].

We utilized the Keras TensorFlow library to build our LSTM model and employed RandomSearch for hyperparameter tuning. This approach is supported by many studies and is considered state-of-the-art [8]–[10]. We also penalized the model by using L2 Regularization. This helps to address over-fitting issues. The code snippet 1 illustrates our approach to obtaining the best model:

In addition to the LSTM model, we also constructed a logistic regression model using the sklearn library. For parameter tuning, we used GridSearchCV with 5-fold cross-validation. This is a common practice [11]. Logistic regression has been used in various financial contexts, including factors affecting financial knowledge in decision-making8 and stock market prediction [11], [12].

Our dataset was divided into training, validation, and test

```
def get_best_model(self):
    tuner = RandomSearch(
        self.build_model,
        objective=Objective("auc",
        direction="max"),
        max_trials=10,
        executions_per_trial=2)

    tuner.search(self.X_train,
    self.y_train,
                 epochs=5,
                 batch_size=128,
                 validation_data=
                 (self.X_test,
                 self.y_test))

    best_model = tuner.get_best_models(
    num_models=1)[0]
    self.model = best_model
    return best_model
```

Figure 1: *Hyperparameter Tuning Codeblock*

sets. We reserved 33% of the data as test data, which was not seen by the model during training. The remaining data was further split into training and validation sets in an 80/20 ratio. The validation data was used to improve the model over epochs when training the Keras model. We trained our models for 20 epochs. Figure 2 shows the evolution of the model over the epochs.

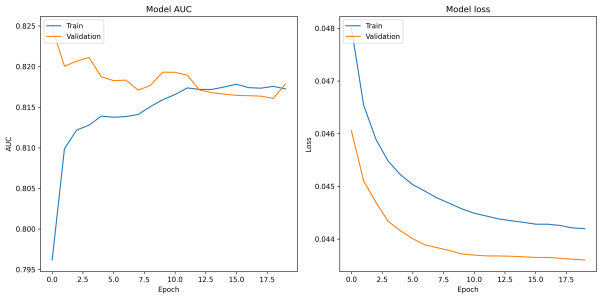

Figure 2: *LSTM Training Performance*

3.1 Application

In our project, we have developed a user-friendly interface leveraging data science methodologies to enhance predictive modeling. The interface, built using Streamlit, facilitates the recommendation of foreign countries to a specified company for a particular year.

These models are trained on historical data to provide accurate forecasts. Upon user input, our system tests each country within the dataset against the given company and year, ranking them based on probability scores. The top five countries with the highest probability scores are then presented as recommendations.

The interface offers a seamless user experience with intuitive inputs. Users begin by selecting the target year for

operation. Subsequently, they specify the home country of the company. Finally, users choose the company from a comprehensive list generated based on the provided criteria.

4 Evaluation

In the evaluation of our predictive models for recommending foreign countries to businesses, we encountered a significant class imbalance within our dataset. Out of 366 413 instances, only 3 267 were labeled as true positives (1), while the majority were labeled as negatives (0). This imbalance posed a challenge for traditional classification approaches, as our models tended to predict the majority class (0) most of the time. Even when employing a commonly used cutoff value of 0.5 for classification, our models consistently predicted 0, rendering the classification ineffective for our purposes.

To address this issue, we adopted a different approach that focused on utilizing the response values of our models rather than simply classifying instances as 0 or 1. Instead of relying on binary classification, we leveraged the probability scores generated by our models to rank potential foreign countries. By doing so, we aimed to provide more nuanced and actionable recommendations to businesses seeking international expansion opportunities.

In the evaluation phase, we conducted quantitative analysis and performance assessment by computing the Area Under the Curve (AUC) scores for our models. Both the Long Short-Term Memory (LSTM) and Logistic Regression models demonstrated noteworthy AUC scores of 0.84, underscoring their efficacy in discerning between positive and negative instances. Detailed metrics are presented in Table 1. It's imperative to note that our evaluation primarily focuses on classification scores. Given the significant class imbalance previously discussed, the models predominantly predict 0 due to the response values falling below the 0.5 cutoff threshold.

To gain deeper insights into model performance, we investigated the predicted investment recommendations. We prioritized responses based on their values ranging between 0 and 1, assessing whether the country of investment ranked within the top 10 recommendations. For instance, when examining Siemens' investment in China in 2014, we input the corresponding company and country data for that year and extracted recommendations from our models. We then assessed whether China featured among the top 10 recommended countries. It's worth mentioning that investment decisions often involve subjective considerations rather than adhering strictly to objective benchmarks. Hence, our model may suggest countries better aligned with the investment opportunities than those actually pursued by the company.

To enhance computational efficiency, we sampled only 50% of the positively labeled investment choices from our dataset. Subsequently, we extracted the top 10 recommendations and determined the percentage of instances where the actual country was included in this list. Approximately 47% of predictions included the country within the top 10 recommendations, with the logistic regression model yielding a slightly lower figure of 46%. When narrowing the focus to the top 5 recommendations, the country was included in approximately 30% and 31.5% of predictions for the LSTM and Logistic Regression models, respectively. These findings suggest that both models identified different investment opportunities compared to those pursued by the company in the majority of cases. Assessing whether the top recommendation outperformed the chosen investment is a complex task and warrants further investigation. It's crucial to acknowledge that machine learning models operate on intricate pattern recognitions, which may not be readily interpretable compared to more conventional decision-making approaches. This underscores the importance of future research endeavors in this domain.

Moving to the qualitative evaluation, we acknowledge that the models predominantly predict 0, and the recommended choices occasionally diverge from the investments actually undertaken by the company.

Model	Recall	Precision	F1-Score	AUC
LSTM Network	0.99	0.98	0.99	0.8422
Logistic Regression	0.99	0.98	0.99	0.8388

Table 1: *Quantitative Model Performance*

Figure 3 elucidates the top 10 most influential features that govern the Logistic Regression model's predictions. Notably, among these features, the Eigenvector centralities of source and destination countries within the network of International Governmental Organizations (e.g., UN) emerge as pivotal determinants.

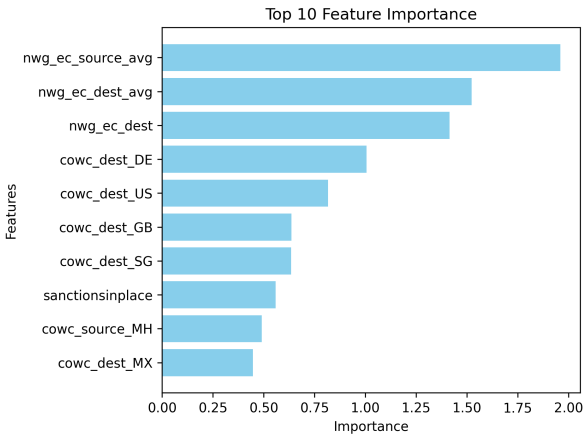

Figure 3: *Feature Importance*

In the realm of International Governmental Organizations, each country functions as a node within a network, with connections representing memberships or participations in these organizations. The Eigenvector centrality of a country in this network serves as a metric of its influence or prestige on the global stage. Countries with higher Eigenvector centrality not only possess extensive memberships but are also connected to other countries with similarly high centrality scores, indicating a significant presence in international affairs.

We can also observe that investments directed towards countries like Germany, the United States, and the United Kingdom exhibit higher likelihoods compared to investments

in other nations such as Hungary. This discrepancy can be attributed to the prevalence of True Positives associated with these countries within our dataset. Moreover, we found that the presence of sanctions between the source and destination countries significantly impacts the model, expediting decision-making processes.

In an illustrative scenario, we conduct a detailed analysis to evaluate and interpret the recommendations provided by our predictive models. Focusing on investment recommendations for a German company specializing in Packaging Machinery in the year 2018, we select Krones AG (ISIN: DE0006335003) as our target company for investment analysis. Table 2 displays the top 5 recommendations generated by our models for this search request.

Suggestion	LSTM Network	Logistic Regression
1	FR	US
2	US	GB
3	CN	FR
4	NL	IT
5	GB	ES

Table 2: *Top 5 Recommendations*

Interestingly, we observe a deviation in recommendations compared to historical investment patterns. Notably, in 2015, Krones AG invested in Columbia, yet our models do not recommend this country for investment in 2018. Instead, the LSTM model suggests France, while the Logistic Regression model favors the United States. This variance is intriguing, especially considering that historical data indicates a higher propensity for German investments in the United States and the United Kingdom, with France not featuring prominently in past investment trends.

Figure 3 illustrates that the Logistic Regression model heavily relies on country connection data, prioritizing historical investment patterns and country relationships. Conversely, the LSTM network appears to place more emphasis on company-specific information, indicating a nuanced understanding of individual company dynamics in its ranking process.

Expanding our analysis to include similar companies based on dataset statistics, we utilize cosine similarity to identify comparable entities to Krones AG. Notable candidates include Xano Industrier (Sweden), Signature Aviation (United Kingdom), and Multi-Color Corporation (United States). Interestingly, the model responses exhibit similarity for companies, with France appearing in the top 5 recommendations. This suggests that recommendations may extend beyond similar countries to encompass comparable companies.

However, despite similarities in evaluation metrics, the models yield slightly different results. The LSTM network leverages company-specific data to a greater extent, yielding more personalized recommendations. However, its lack of explainability may be a drawback. In contrast, the logistic regression model offers greater interpretability. Therefore, the choice between models depends on the importance of explainability in the given context.

5 Conclusion

In conclusion, our study leveraged extensive datasets sourced from Orbis Crossborder, FDIMarkets, and Refinitiv, meticulously curated and transformed by Dr. Wolfesberger. These datasets provided a comprehensive repository of foreign subsidiaries worldwide, enabling our analysis of investment opportunities undertaken by companies across various industries.

Two distinct datasets were employed: one focusing on company-specific attributes and investment opportunities, while the other provided insights into country connections, including trade agreements, conflicts, and cultural distance.

Our approach involved building two predictive models: a Long Short-Term Memory (LSTM) model developed using the Keras TensorFlow library, and a logistic regression model implemented using the sklearn library. Hyperparameter tuning for the LSTM model was achieved through RandomSearch, supported by L2 Regularization to mitigate overfitting. Similarly, for the logistic regression model, parameter tuning was conducted using GridSearchCV with 5-fold cross-validation, following common practices in the field.

Our analysis yielded insightful recommendations for investment opportunities, considering both historical trends and company-specific factors. Notably, the LSTM model demonstrated a nuanced understanding of individual company dynamics, while the logistic regression model relied more heavily on historical country connections and average investment trends.

References

[1] S. S. Roy, "Prediction of foreign direct investment: An application of long short-term memory," *Research Gate*, 2020.

[2] C. S. R. Yu Y. Li, "Deep learning: A generic approach for extreme condition traffic forecasting.," *SIAM International Conference on Data Mining*, pp. 777–785, 2017.

[3] G. R. Zimmermann H. Tietz C., "Forecasting with recurrent neural networks: 12 tricks. in: Neural networks: Tricks of the trade.," *Springer*, pp. 687–707, 2012.

[4] E. C. Lipton C. Kale C., "Learning to diagnose with lstm recurrent neural networks.," 2015.

[5] M. Wolfesberger, "Configuring and restructuring mncs' foreign subsidiary portfolios in a challenging political environment," 2023.

[6] M. C. Cheng Fan, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers*, 2021.

[7] H. Q. André Carrington Paul Fieguth, "A new concordant partial auc and partial c statistic for imbalanced data in the evaluation of machine learning algorithms," *Springer*, 2020.

[8] M. He, "Deep learning for dynamic nft valuation," *Arxiv*, 2023.

[9] L. Tao, "Unified machine learning protocol for copolymer structure-property predictions," *ScienceDirect*, 2022.

[10] L. Biedebach, "Anomaly detection in sleep: Detecting mouth breathing in children," *Springer*, 2023.

[11] I. M. Isfenti Sadalia Fahmi Natigor Nasution, "Logistic regression analysis to know the factors affecting the finan-

cial knowledge in decision of investment non riil assets at university investment gallery," *SSRN*, 2020.

[12] C. L. Wessel van Eeden, "Predicting the 9-year course of mood and anxiety disorders with automated machine learning: A comparison between auto-sklearn, naïve bayes classifier, and traditional logistic regression," *ScienceDirect*, 2021.