National College of Ireland

Using Machine Learning Techniques for optimizing Research and Development in Irish Businesses

MSc Research Project
Data Analytics

Conor Moody
Student ID: 21201765

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Conor Moody
Student ID:	21201765
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	14/08/2023
Project Title:	Using Machine Learning Techniques for optimizing Research and Development in Irish Businesses
Word Count:	7,908
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Conor Moody
Date:	18th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Using Machine Learning Techniques for optimizing Research and Development in Irish Businesses

Conor Moody
21201765

Abstract

Machine Learning is increasingly allowing businesses to gain greater insight into the metrics that impact their performance. This project looked at the Annual Business Survey of Economic Impact (ABSEI), which is distributed to thousands of businesses throughout Ireland. The relative obscurity of this dataset in the academic field posed challenges and opportunities for exploratory analysis. By looking at different implementation methods, this project attempted to estimate the appropriate amount of spend on Research & Development that a company should invest in. As the dataset contained data spread over 23 years, the time series potential of a Recurrent Neural Network with Long Short-Term Memory was utilised. A Feed Forward Neural Network was also looked at due to the regression elements of the project. Finally, a Linear Regression model completed the comparisons, and the 3 models achieved a good fit, allowing for a greater understanding of the dataset that will better serve Irish businesses in the future.

1 Introduction

Research & Development (R&D) is a vitally important part of many businesses. Investing in R&D is one of the biggest risks that companies can take. This research project will look towards the advancements in machine learning to assist with mitigating that risk.

The Annual Business Survey of Economic Impact (ABSEI) has been recording data on Irish businesses for the past 20 years. It is coordinated by the Department of Enterprise, Trade & Employment. More than 7500 Irish businesses have contributed to the survey, and the data gathered allows for greater insights into how these businesses operate. The focus of this project is to determine how this data can be leveraged to promote growth for these companies. In particular, machine learning tools were implemented to predict the amount of expenditure a business should invest in R&D.

The application of Machine Learning tools in a business environment is a relatively new concept. In 2020, Reis et al. (2020) commented that it had not yet been substantiated how organizations can derive business value from Machine Learning. This research project will address this problem, with a view to R&D and it's ability to transform the fortunes of a business. There is strong potential in this project for Irish businesses, aligned with the fact that this dataset would not normally be available. Due to the innovative and unique nature of the dataset, there is great value to be had in the results for Irish businesses.

1.1 Research Questions

In order to best demonstrate the ability for machine learning to facilitate investment decisions on R&D for Irish companies, this project posed the following questions.

Main Research Question : *To what extent can machine learning tools recommend the value of Research and Development for Irish companies?*

Sub-Research Question 1 : *How can a businesses ability to invest in Research and Development be affected by exports?*

Sub-Research Question 2 : *To what extent does the cost of energy impact investment in Research & Development?*

1.2 Research Objectives

R&D can generally fall into 3 different categories. These are Basic Research, Applied Research, and Development. Basic Research attempts to understand a subject matter in more detail, but it does not bring too much practical or commercial application. Applied Research is a more targeted approach and looks towards addressing a commercial need. Development is when the results of research are then utilized to produce specific products. Depending on the category that R&D falls into, a time lag of varying length needs to be factored into building any model that challenges the impact of R&D on new product sales.

Another important factor at play is the fact that some firms will engage in persistent R&D, while others will only engage in intermittent R&D. Persistent R&D is usually the only beneficial method for smaller companies, whereas larger companies can derive benefits from both methods.

The objectives of this research will be to create a machine learning model capable of predicting a value for R&D for the businesses in the dataset. This will allow similar businesses to more accurately structure their expenditure, and will assist with future submissions to the survey.

As this dataset is not widely known in the academic field, there is a large exploratory nature to the work being carried out. The need for multiple models was important in order to know how best to use the data.

1.3 Structure of Report

This introduction sets out the background to this project. It is important to look at the related work on this topic in the academic fields. This frames the decisions that follow for the Implementation phase. The Methodology section looks at the acquisition of the data as well as the theory behind the work to be carried out. The Design Specification will discuss the library that was chosen as well as the choices taken when designing the model. From here, the implementation of the model is discussed and then the results are evaluated in order to answer the posed research questions. The conclusion will summarize the work done and pose the potential for future work in this area.

2 Related Work

2.1 Introduction

The dataset utilised for this project was unique and innovative in its design. However, there was much knowledge to be gathered from papers in the academic field in terms of solving the posed research questions. In particular, the application of chosen machine learning techniques was of great interest. This research project looked at papers from the previous 5 years, as Machine Learning in business planning is a relatively recent occurrence. By assessing related work, there was a focus as well on the sub research questions and the area of study around them.
In looking at Related Work in this area, a generalised view will be taken of Machine Learning in business environments. Particular focus will then be given to the area of R&D, in the context of the subject matter of the main Research Question. Neural Networks are an important part of this research project, and their impact in related works will be discussed in this section. Other factors relevant to business planning, such as Exports and the impact of Climate Change will be looked at with a view to the sub research questions posed in this paper.

2.2 Machine Learning in a Business Environment

With any model attempting to facilitate business planning, a number of important factors need to be considered. Some models may work better for other areas. There are also extraordinary factors that may occur, causing major disparity with past data. The COVID pandemic is an example of this. This was looked at by OTrakoun (2022) when they observed that for the automotive industry and goods manufacturing in the information sector, forecasting strategies that were projection based were found to be more resilient than iteration-based forecasts, in the aftermath of the COVID pandemic.
Productivity increases of 5% and profit increases of 6% were observed by Reis et al. (2020) as a result of incorporating data and analytics in a company. While also assessing the drivers of machine learning business value, they found that platform maturity has a catalytic effect on business value. However, as previously mentioned, there are issues with looking at data in a moment of time, especially when geo-political and global health factors have a major impact on the planning process. The data gathered for this research project was longitudinal, and this posed the potential for it to provide stronger insight as a result. The application of longitudinal data in a machine learning model was something achieved by Cao et al. (2019). They implemented a time series forecasting model based on Long Short-Term Memory (LSTM). They reduced the impact of noise on the prediction by using empirical mode decomposition (EMD) and complete empirical model decomposition with adaptive noise (CEEMDAN). This might not be as relevant for predictions of non-financial data, but their use of stock price data over a 10-year period draws similarities to the dataset employed in this research project. In particular, the LSTM approach to dealing with time series data was of potential use.
With a variety of different attributes available for businesses when looking to incorporate machine learning models into their planning, it is important to avoid unstructured noisy data. When looking at this, Angenent et al. (2020) used a Random Forest classifier on over 1.5 million anonymous financial statements. They discovered the top occurring features, and stated their value as a measure of importance. For this they observed that business sectors are predominantly characterised by a small subset of attributes. A Fuzzy

Chance Constrained Least Squares Twin Support Vector Machine (FCC-LSTSVM) was employed by gang Song et al. (2018) when also looking at financial data. They attempted to predict business performance through financial ratios, using 796 companies in China. Their results echoed previous findings that indicated that there were different achievements for predicting business performance depending on the particular industry. The need for business diversification has lead to this stronger focus on machine learning. However, businesses have been too focused on looking at technological capabilities instead of applying scientific methods to historical data (Lee et al. 2020). By looking at historical data, the benefit of time-series data can be realised. This was implemented by Jamshed et al. (2020) whilst primarily looking at sequential pattern mining. By designing a hybrid convolutional neural network (CNN) with long short-term memory (LSTM), they attempted to discover customer behavior and purchasing patterns for online businesses. In general, the ability for a business to streamline its processes will introduce greater opportunities for investment in areas such as R&D. Park & Song (2020) looked at the modern ability to not just analyse business processes based on historical data, but to look at operational data to make real-time decisions that mitigate business risks. They showed how their methods based on deep neural networks outperformed baseline approaches, "by successfully learning the temporal evolution and the spatial dependency.

2.3 Machine Learning Applications for Research & Development

This research project will look at proposing the optimal spend as a % that businesses should adopt in order to maximise growth. The impact of higher R&D spends were observed by Lee et al. (2019) when they attempted to determine if more investment in R&D would lead to improved company performance and whether increased R&D expenses were a worthwhile investment. Using a Backwards Propagation Neural Network to assist in assessing the importance of the use of R&D, they were able to determine that R&D intensity and operational efficiency play a pivotal role in helping companies attain better capabilities in boosting their market performance. One key observation made by Lee et al. (2019) was that in mainstream manufacturing, R&D has been consistently considered a critical strategic determinant of future growth. They also noted that increases in R&D spend led to market value increases as it reflected positively on investor optimism.

2.4 Application of Neural Networks in a Business Environment

Neural Networks are the most prominent applications of Machine Learning in Business Environments. Zhang (2022) built a neural network using random matrix theory. This was done with a view to countries needing to have competitively stronger advantage in the international market. It was found that Random Matrix Theory was more accurate than a hierarchical analysis method in evaluating business performances in enterprises, particularly corporate performance.

In general, the ability to forecast was key to answering the research questions posed in this paper. By utilizing a Bayesian Belief method, Kapoor & Wilde (2023) obtained higher accuracy results than a standard Bayesian forecasting method, and highlighted that you can achieve superior industry foresight where individuals update their beliefs in learning-based forecasting behaviour and then weigh those initial beliefs carefully with new information. A backwards propagation neural network was employed by Li (2023)

to predict economic growth in a Chinese province. They optimized the neural network using a genetic algorithm, and helped to reduce the relative error in models while greatly increasing forecast precision.
An issue that is not addressed by Kapoor & Wilde (2023) is the nonlinear attributes of a macroeconomic system. Due to variables such as a recession or the pandemic, it is a constantly changing environment that limits the reliability of traditional forecasting. This issue, and the common issues with normal back propagation neural networks were addressed by Liu et al. (2022). They proposed to solve these by implementing a Recurrent Neural Network, with Particle Swarm Optimization on Gated Recurrent Units (PSO-GRU). This helped them achieve higher accuracy that standard GRU or Long Short-Term Memory (LSTM) models as well.
Neural Networks are also commonly used in Recommender Systems. The ability to look at long term data was an important aspect of the dataset used in this project. When looking at the implementation of neural networks in recommender systems, Bansal & Baliyan (2022) also saw greater accuracy with a GRU than with a LTSM model. Different companies may require different needs from a recommender system, and machine learning can assist with this. With a view to the impact of COVID-19 on households, a recommender system was designed by Kannan et al. (2021) to analyse the financial assistance needs of households and to create a customized economic stimulus package for them.

2.5 Factors Influencing Research & Development Spend

As stated in the Introduction, there are different types of R&D and these will take varying amounts of time to bring a return. There is also the decisions for businesses to continue investing in R&D each year or to sporadically invest at particular intervals. Holl (2021) looked at this issue in the context of business in Spain. They found that "a stronger regional knowledge environment increases the likelihood that an innovator engages continuously in R&D as opposed to occasionally".
The ability for a company to invest in R&D is dependent on a number of factors in terms of expenditure. One of these in particular is the increasing need for businesses to adapt to climate change. There is a lack of awareness of this by way of corporate adaptation, as noted by Pinkse & Gasbarro (2019). Drawing on the attention-based view (ABV), they highlight that companies adaptation measures are generally routine measures, as opposed to the non-routine aspects of climate events.
A factor not considered by Pinkse & Gasbarro (2019) was the ability for Small to Medium Enterprises (SMEs) to adapt to climate change. When looking at this Fawcett & Hampton (2020) made light of the fact that SMEs uses half of business energy in the UK and the EU. So any policy changes mandating businesses to adjust their energy usage needs to reflect this. The ability for SMEs to invest in R&D is closely linked to this as well.
These policy changes will be unwelcome for a lot of businesses. In an observation of business operating from Hunter Valley in Australia, Forino & von Meding (2021) highlighted than some businesses are sceptical of climate change, while others are cognisant of the risks to themselves and relevant stakeholders of a failure to adapt. They argue that further research is needed to fully judge the impact of climate change on a businesses expenditure.
The economic recession that began in 2009 caused a big impact on the ability for companies in Europe to invest in R&D. Focusing primarily on Germany, Belitz (2022) note

that R&D spend decreased initially in the first few years following the recession, and then steadily increased afterwards. However, there was a decrease observed in 2020. This decrease highlighted that certain sectors are spending more on R&D than others. These are sectors that Germany does not specialize in, such as software and computer services, hard-ware production, as well as pharmaceuticals and biotechnology.

2.6 Conclusion

By observing the related work in this field, there are a number of important observations that can be derived. There is an emerging need for machine learning algorithms to bring significant predictive benefits for businesses looking to carefully position their expenditure to maximise income. The role of data science in economics applications is something also looked at by Nosratabadi et al. (2020). Their observations point towards a sector that experienced a 4-fold increase in the applications of data science in economics between 2016 and 2019.
Another important observation, also seen by Nosratabadi et al. (2020) is the emergence of hybrid neural networks. It's been proven in this section that hybrid models have obtained greater accuracy than base neural network models. This was seen by Liu et al. (2022), Bansal & Baliyan (2022), Li (2023) and Zhang (2022) among others. Assessing company data can be done in a number of ways. Stichhauerova et al. (2020) looked at the performance of clustering on companies, achieving statistically significant p values (less than 0.05) for each cluster. However, the ability to predict a value for R&D carries more long term benefits, and the importance of comparing models is seen by Hrnjica & Bonacci (2019) when they looked at predicting lake levels using Feed Forward and Recurrent Neural Networks. They showed that both models gave satisfactory results compared to an ARIMA time series model.

3 Methodology

In order to solve the research questions posed by this project, the related work set guidelines for the methodology to follow. The result of that research pointed in the direction of a Recurrent Neural Network (RNN) to make best use of the time series element of the dataset. The model should also apply long short-term memory (LSTM) in order to combat the vanishing gradient problem inherent in recurrent neural networks. With a view towards the exploratory nature of this project, there was a strong need to implement alternative models that made better use of the data. The regression elements of attempting to predict a value for R&D suggested that a Feed Forward Neural Network would make a good comparison to the RNN LSTM model. It was also useful to examine whether a simpler model may achieve the same results with less complexity and processing intensity. As such, a Linear Regression model was implemented on the dataset as well.

The methodology outlined in Figure 1 describes the steps in the process in order to build the models required to answer the research problem.

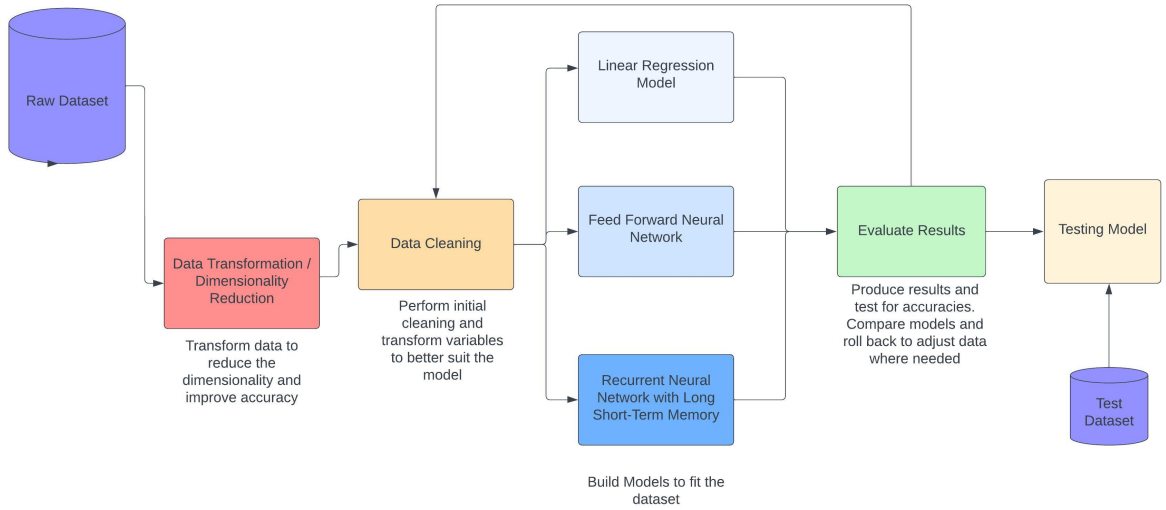

Figure 1: Methodology

Once the data was transformed, it needed to be refined in order to best suit the models that will be built on it. This is explained in more detail in section 5.1. It was important that all 3 models were built on the same cleaned data, to provide the most effective comparisons between each of them. Once evaluated, the models could be deployed for use with test data.

3.1 Neural Networks

Neural Networks are made up of layers of individual units called neurons or nodes, illustrated in Figure 2 below, which attempt to mimic the biological behaviour of the brain.

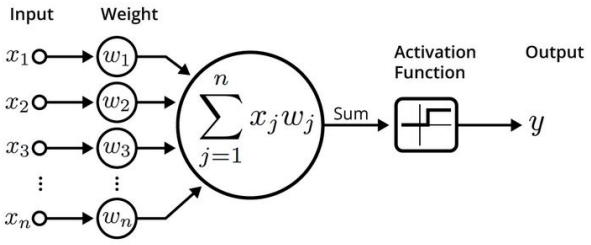

Figure 2: Neuron in Neural Network

Each node can consist of inputs (from the dataset or from a previous node), weights (to give importance to features that contribute more to learning), transfer functions (to combine multiple inputs into one output for the activation function), and the activation function (which calculates the output for the node).

3.2 Data Acquisition

The data from the Annual Business Survey of Economic Impact (ABSEI) is generally available to researchers on request. However, particular access to restricted parts of the survey was permitted for this project. As the ABSEI survey has generally focused on income, expenditure and exports, many of the R&D elements of the data have not been previously available to other researchers. Approval was given to this research project to access a version of the dataset containing the information on R&D. The dataset contains the information on 10,856 companies. These companies operate in Ireland, but many are multinational whose country of origin is not Ireland. For this project, the focus was on the 7,803 Republic of Ireland companies who returned results in the survey over the 23 years that the data was gathered.

3.3 Data Transformation

When the data was received, it was clear that it was a high dimensional dataset, in that there were a large number of variables in the dataset compared to the number of records. To overcome this obstacle, dimensionality reduction was performed on the dataset to reduce the number of features and to lower its high dimensional aspect.
The survey data for this project recorded individual entries for each question and for each year the question was asked. This meant that every question in the survey could contain 23 variables in the dataset, if that question had been asked each year from 2000 to 2022 inclusive. The survey contains 185 questions, and although not all questions have been in the survey since 2000, this still amounted to a dataset of 1059 variables. Given the number of companies in the dataset, this posed a dimensionality issue that would cause issues with the analysis. A dataset with high dimensionality (a high number of features in relation to observations) can create problems in machine learning as large training data sets would be needed to best represent the likely combination of values. Computational Complexity as well as overfitting can also be issues arising from high dimensionality.
The dataset in its original state needed to be transformed in order to lower the dimensionality and improve the performance of the model. Another major reason for transforming the data was to allow for a proper use of the time series elements. It was not possible to perform time series analysis on the dataset in its current state.
In order to transform the data, the 'melt' function in Python was implemented on those variables which carried multiple years' of answers. The melt function is used to reshape a DataFrame into a format where one or possibly more columns are identifier variables, while the other columns, which are considered measured variables, are un-pivoted to the row axis. In this dataset, this meant creating a new row in the dataset for every year from 2000 to 2022. The base variables for each company were then populated into those rows, while also then having just one variable per question in the survey, and inputting the companies response for each year in the accompanying rows. Table 1 shows how the data was originally presented, and Table 2 shows the data after it was reshaped.

Table 1: Original Format

Company ID	Origin	x1 2000	x1 20001	x1 2002
1000001	Ireland	4500	5500	4250
1000002	United Kingdom	1950	2850	2400
1000003	Ireland	8900	7800	7450

Table 2: Reshaped Format

Company ID	Year	Origin	x1
1000001	2000	Ireland	4500
1000001	2001	Ireland	5500
1000001	2002	Ireland	4250
1000002	2000	United Kingdom	1950
1000002	2001	United Kingdom	2850
1000002	2002	United Kingdom	2400
1000003	2000	Ireland	8900
1000003	2001	Ireland	7800
1000003	2002	Ireland	7450

The dataset has now been transformed from 1,059 variables and 10,856 observations into a dataset with 185 variables and 249,664 observations. There is now an observation for each company, for every year that they submitted a response to the survey. This makes the dataset far easier to use from a time series point of view and has lowered the dimensionality of the overall data. As a result of this work, no other dimensionality reduction techniques were carried out on the dataset.
Tabulating the data was also an important element to the data preparation. By utilizing the tabulate function in Python, the dataset has been moved into a simplified format and it establishes relationships between the variables. It also makes the data easier to comprehend for analysing.

3.4 Exploratory Data Analysis

The dataset contains companies operating out of Ireland. However, there are a large number of multinationals who do not have their origin in Ireland. As Illustrated in Figure 3 below, 71.9% of companies are from Ireland, with United States representing the largest of the rest of the companies with 13.6%. This project will focus on companies operating and originating from Ireland.

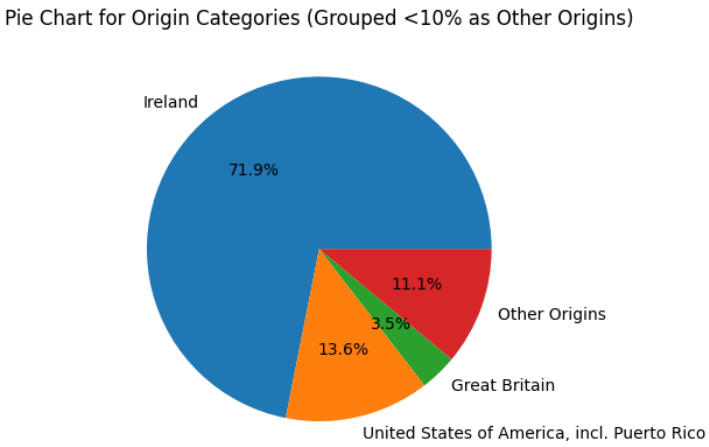

Figure 3: Company Origins

The survey underlying this dataset has been in circulation since 2000. The amount of money spent by companies on R&D has been recorded since then. It is recorded in the dataset as the variable x25, with the value recorded in the 000's. Figure 4 below shows the distribution of that spend in the lifetime of the survey.

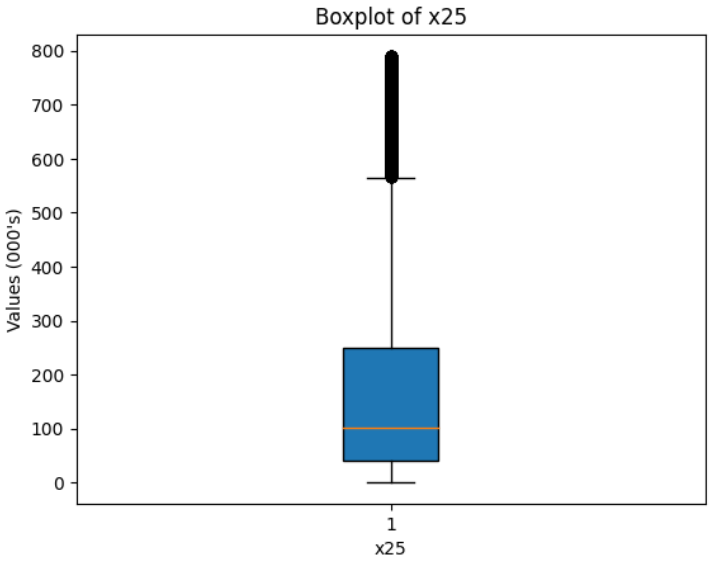

Figure 4: Distribution of R&D spend across the dataset

Figure 5 shows this distribution across each year, which shows a small but steady increase in the investment in R&D over a 22 year period.

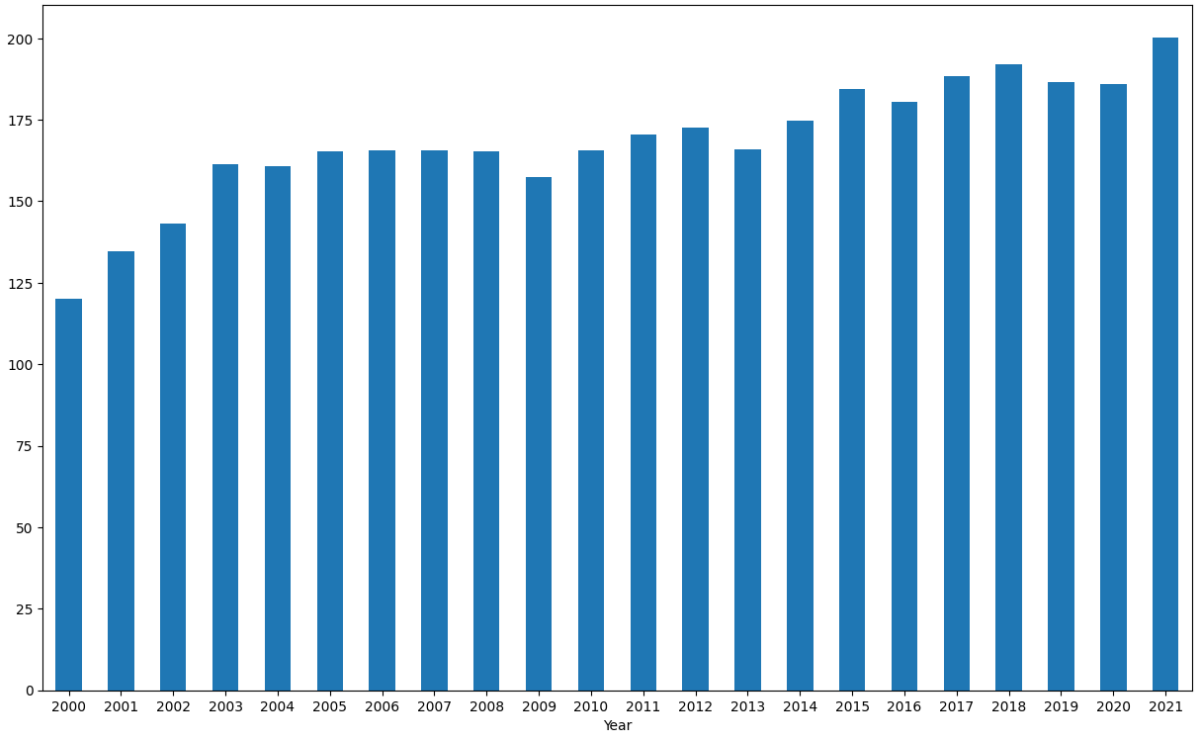

Figure 5: Average spend (€000's) for companies investing in R&D

3.5 Data Limitations

The main dataset carries values over a 23 year period, ranging from 2000 to 2022 inclusive. However, questions are R&D only cover the years up to 2021. Also, the questions in the survey around New Product Sales have only been present since 2012. This gave just 10 years of this data, and presented challenges for correlations between R&D spend and the translation of that spend into New Product Sales. As discussed, there is also no data on the type of R&D that these companies carried out. This means that it is not clear whether the R&D carried out by these companies was basic research, applied research or development. As such, the translation of this work into new product sales was difficult to quantify. A calculation was made during the Implementation phase to translate R&D expenditure into New Product Sales was, and is discussed further on page 13

3.6 Feature Selection

After the transformation phase, there are 185 variables in the dataset. These are reflective of the answers to the ABSEI survey for the 23 years that the survey has been conducted. However, as has been stated, not all of questions were asked in every year. In particular, there are 97 variables out of the 185 where there is only data for 2021. In the data preparation phase, these variables were removed. In order to combat the impact of outliers on the model, any records containing an outlier for the x25 variable were also removed.

4 Design Specification

The design of this research project followed the layered architecture approach to software development. Figure 6 below illustrates this architecture.

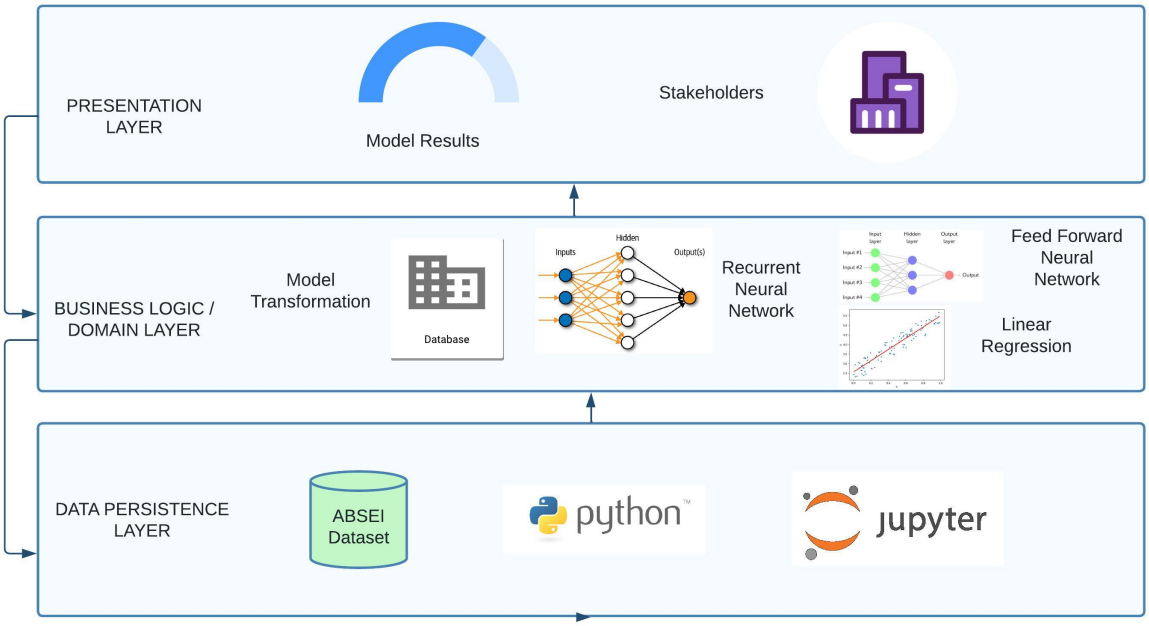

Figure 6: Project Design

4.1 Chosen Libraries

At the root of the design are the chosen tools that will allow for the development of the Recurrent Neural Network model. The Python programming language was used to perform the work. It was chosen for this project as it has a number of useful libraries that can be used to perform the required actions. Jupyter Notebook is also a very useful development and visualisation tool, and was used alongside Python to enhance the output.

4.2 Model Design and Stakeholders

The Recurrent Neural Network sits at the heart of the design, and other algorithms were implemented to attempt to improve accuracy and assist with the answering of the posed research questions. The results of the model were analysed with the goal to maximise the impact for the Irish companies who submit their data to the survey. Any transformation that was done to improve performance was fed back down to the domain layer.

5 Implementation

5.1 Preparation

Following the Data Transformation, there was still a lot of work needed to prepare the data for modelling. As this dataset had not widely been used academically, the work to prepare it was not able to rely on best practice for dealing with some of the complexities of it.
As this project looked at Irish companies, it was important to remove any companies whose origin was outside of the Republic of Ireland. From an economic point of view, this provided the most stable ground for building an effective model to predict the appropriate spend on R&D.
Also, any data containing strings needed to be removed. These removed a number of variables from the dataset, bringing the number down to 128.
In order to limit the effect of outliers on the performance of the models, a decision was taken to remove any outliers from the x25 variable where the value was 1.5 * Interquartile Range (IQR) above Q3 (3rd quartile) or below Q1 (1st quartile).
One of the main benefits of performing R&D is to allow for a company to create new products as a result of the process. The dataset contains a value for New Product Sales, but as has been discussed, it was not known what type of R&D the company carried out. As such, there was a strong need to prepare the data with the correct realization of New Product Sales values for each year that R&D was carried out A formula was built to see when the highest value of New Product Sales was recorded, after R&D had been spent in a given year. The result of these calculations showed that on average, 5.4 years was needed to realize the results of R&D spend. The value for both x38a ('Total New Product Sales') and x38b ('% New Product Sales') were then amended in the dataset so that for a given year, the value for x38a and x38b in that year were replaced with the value for x38a and x38b from 5 years later. So as an example, the data from 2006 contains the New Product Sales figures for 2011.

5.2 Feed-Forward Neural Network

The layer architecture of a Feed-Forward Neural Network (FFNN) is similar to the architecture of other neural networks. It is composed of an input (or inputs) layer, hidden layers and an output layer. The architecture in Figure 7 represents a multi-layer Feed-Forward Neural Network.

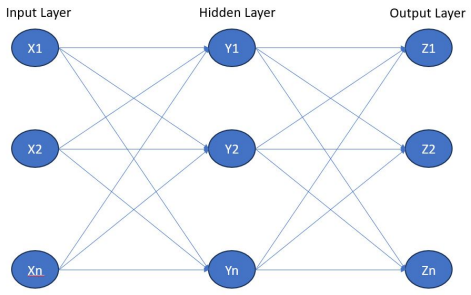

Figure 7: Feed-Forward Neural Network Architecture

The first stage in the implementation of a FFNN on the dataset was to normalize the

data. Carrying this out allowed for all the features to be on the same scale as x25. This ensured the FFNN learned more effectively. It also allowed for each feature to contribute equally to the predictions, and improved the performance overall.
Once x25 was removed from the dataset and assigned to the variable 'y', the dataset was then split with 80% being used to train the model, and the remaining 20% left to test it.The input features were then standardized to mitigate against issues such as vanishing or exploding gradients. These issues were more likely to occur in the RNN LSTM model, but it was chosen given the size of the dataset and the type of functions being used in creating the model. This standardization also helped to protect against overfitting of the model.
The FFNN model was then created using a Multi-Layer Perceptron Regressor (MLPRegressor) function. This function is designed for this type of a regression task, where we're looking to predict numerous x25 values. MLPRegressor is easily customized and it supports many different activation functions and optimization algorithms. This allowed for it to be changed to better optimize the model. The Rectified Linear Unit (ReLU) was then used as the activation function. It was chosen as it is a popular activation function for feed forward neural networks. It can help bring non-linearity into a network, making it easier to learn relationships and patterns in the data. The Adaptive Moment Estimation (adam) solver was used to optimize the model and reduce the loss function. Adam computes individual learning rates for each parameter, speeding up the convergence process. It incorporates momentum to help with this, and also contains bias correction for the initial estimates of the first and second moments, ensuring accurate estimates during early iterations. It's also robust towards hyperparameter choices. 50 Hidden Layers were found to be most optimal for this model, and a random state parameter of size 42 was used to control for randomness in the training of the model. The model was then trained and predictions were obtained for x25. The results of these are discussed in section 6.1

5.3 Recurrent Neural Network

As the dataset contains company data over a 23 year period, it was important to assess whether a neural network with a time series component could work better for predicting R&D spend in the dataset. As seen in the architecture in Figure 8 below, a Recurrent Neural Netwok (RNN) has connections between nodes that can form cycles, which allows information to persist and flow back into the network. By having this feedback loop it enables RNNs to have a hidden state which is able to capture temporal dependencies in sequential data.

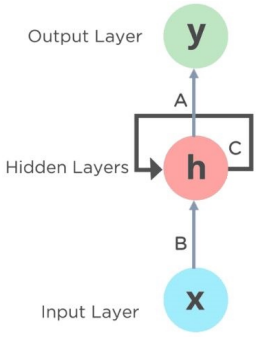

Figure 8: RNN Architecture

One of the issues that RNNs can face is the vanishing gradient problem. This is when the gradients used to update the network become very small to the point of vanishing, as they are back-propagated from the output layers to the earlier layers. In order to overcome the issues of vanishing gradient, Long Short-Term Memory (LSTM) was the type of RNN used for building the models on this dataset. Gated Recurrent Unit was another type discussed in the Literature Review, but it was not chosen as it can't capture long-term dependencies as well as LSTM.
The ability for LSTM models to effectively capture long-term dependencies is illustrated in the architecture in Figure 9. A simple RNN model would just have a single hidden state, making it difficult to interpret long term dependencies. The LSTM model employed here has a memory block containing 3 gates. The Forget gate removes any information no longer needed. The input gate adds useful information to the cell, and the output gate extracts the useful information and sends it as the input to the next cell.

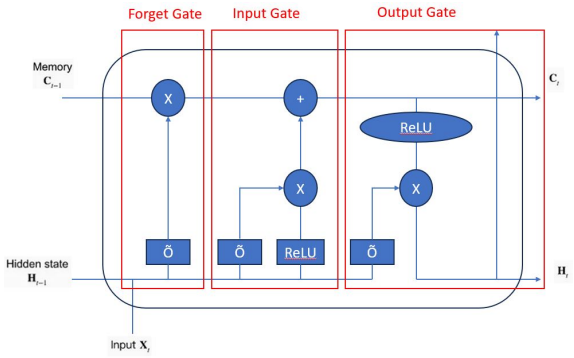

Figure 9: LSTM Architecture

For the implementation of the RNN LSTM model on the dataset, the tensorflow software library was imported into the model. This library allows for training of neural networks and the Sequential, LSTM, Dense and Dropout packages were needed from it to build the model. After a window size was created to determine the number of time steps in the model, sequences of data were created for each unique company (identified by the 'BIS' variable) in the dataset for the specified window size. The data was then split and 'x25' was removed from both the training and test sets.
In order to improve the performance of the model, normalization was carried out on the

data. Min-Max scaling re-scaled the features to be between 0 and 1. This allowed for faster convergence during training. One of the variables in the dataset was a weighting for each year that the survey was carried out. This weight was factored into the model at each stage to apportion the appropriate bias to a given year.
The model was built using the Sequential RNN function, and the ReLU activation function was used over the more conventional 'tanh' or 'sigmoid' functions. This was chosen after hyperparameter tuning as ReLU was found to produce stronger results. Similar to the FFNN model, the adam optimizer was chosen. Dropouts were used to prevent overfitting, and a Dense (fully connected) layer with one neuron for regression output was also included.
When the model was being fit to the data, 100 epochs (to run through the entire dataset) were chosen and 'early stopping' was built into the functionality to prevent overfitting. Early stopping monitors the model's performance on the validation set during the training period and stops the process when the performance begins to degrade, or if the performance does not improve for a number of consecutive epochs. The results of this model are discussed in section 6.2

5.4 Linear Regression

The objective focus of this project was to determine the appropriate spend on R&D in Irish companies. As the nature of these objectives revolved around prediction of values, it was important to implement a Linear Regression model to allow for comparisons with the more complex neural networks being implemented. A simple linear regression model has no hidden layers but looks to fit the independent variables of the model to the target variable. It attempts to find the best fitting straight line to describe the relationship and minimize the sum of squared differences between them.
Using x25 as the predetermined target variable, it was separated from the rest of the data, which was then split into testing and training sets. A model as then built using the LinearRegression function of the sklearn software library. The results of this model are discussed in section 6.3.

5.5 Exports and Energy

The ability for a company to invest in R&D is dependent on a number of factors. The sub questions in this project looked at the how companies can make resources available to invest in R&D, with a particular focus on money taken in from exports, and the cost of energy. The pandemic has caused a shift in the way companies export their goods, and different markets have emerged in the last few years. As has been shown in the Literature Review, climate change is an important factor for many businesses, and will be an increasing influence on expenditure in the future.
In order to assess the impact of these factors on R&D spend, the transformation done on the dataset enabled a correlation heatmap (seen in Figure 10 below) to be implemented to assess the relationship of each variable in the dataset against each other. The darker colours in the map represented stronger correlations. This map was generated after data cleaning, when the number of variables had been reduced to improve efficiency in the model. The impacts of these results on the sub research questions can be seen in section 6.4

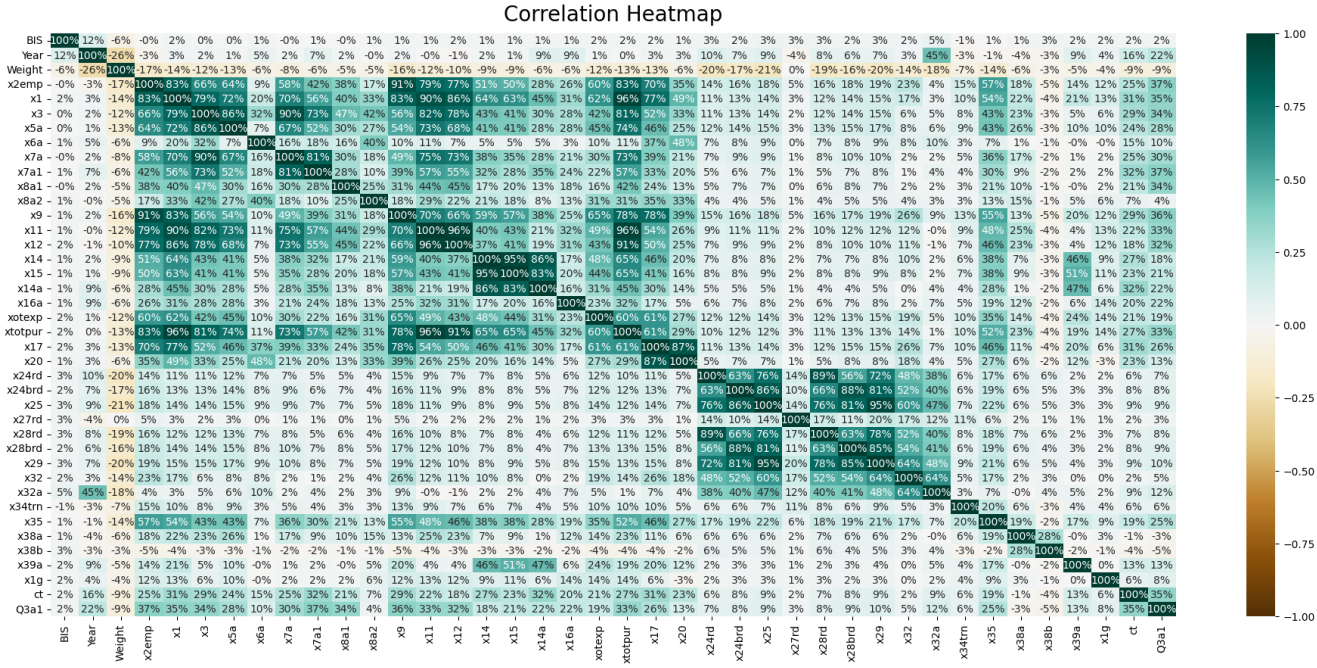

Figure 10: Correlation Heatmap

6 Evaluation

The key objective in evaluating this project is to assess how well the chosen models performed at being able to recommend the investment in R&D in Irish companies. As this was a regression focused task, the evaluation of these results looked at the Mean Absolute Error (MAE), the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE) and the R-squared (R^2) value.
The MAE represents the average of the absolute difference between the actual values and the predicted values in the dataset. As a result, the lower the MAE, the better the model performs. It is represented by the equation below.

$$MAE = \sum_{i=1}^{N} |x_i - y_i| \tag{1}$$

The MSE represents the average of the squared difference between the actual values and the predicted values in the dataset. Again, the lower the MSE, the better the performance. It is represented by the equation below.

$$MSE = \sum_{i=1}^{N} (x_i - y_i)^2 \tag{2}$$

The RMSE is the square root of the mean squared error, so it also seeks a low performance. It is represented by the equation below.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2} \tag{3}$$

R^2 is best described as a goodness of fit test. It is a statistical measure for regression models that determines the proportion of variance in the target variable (x25) that can be explained by the other features in the dataset. The closer the result is to 1, the better the fit of the model.
Table 3 below shows the results for the MAE, MSE, RMSE and R^2 for the three models.

Table 3: Comparison of Results

Model	MSE	MAE	RMSE	R^2
Recurrent Neural Network with Long Short-Term Memory	1,167	20.47	34.1	0.965
Feed Forward Neural Network	1,493	21.3	38.6	0.951
Linear Regression	1,725	22.84	41.5	0.945

The best MSE result was 1,167, as it was the lowest. This result was achieved by the RNN-LSTM model, meaning it had the lowest squared difference between the predicted values and the actual values. The best MAE result was 20.47 which was also achieved by the RNN-LSTM model. This score represented the fact that the model had the lowest absolute difference between the predicted values and the actual values. The RMSE score of 34.1 achieved by the RNN-LSTM model was also the best result. As the RMSE represents the square root of the squared error, it's the result that is on the same scale as the original data. By achieving a score of 34.1, the RNN-LSTM model had the lowest difference between the predicted values and the actual values, of all 3 models. In terms of a fit, the R^2 value of .965 achieved by the RNN-LSTM model was the highest and the best performing, as it was the closet to a perfect fit of 1. The RNN-LSTM model had the lowest MAE, MSE and RMSE, and the highest R^2 value, meaning it performed the best of the three models.

The results of the models can also be viewed in the context of looking at the mean and range of x25. X25 has a mean of 171, and values ranging from 1 to 790. The table below shows the MAE and RMSE performance, viewed alongside the mean and range of x25. This shows how significant these errors are relative to the scale of the data. By comparing the errors to the mean, it's possible to see how large the errors are relative to the average x25 value in the dataset. And by comparing the errors to the range, it's possible to see how large these errors are relative to the overall spread or variability in the dataset. The comparison of these results is illustrated in Table 4 below.

Table 4: X25 Mean & Range Comparisons

Model	MAE (% of x25 Mean)	RMSE (% of x25 Mean)	MAE (% of x25 Range)	RMSE (% of x25 Range)
Recurrent Neural Network with Long Short-Term Memory	11.97	19.94	2.59	4.15
Feed Forward Neural Network	12.45	22.57	2.69	4.86
Linear Regression	13.35	24.26	2.89	5.25

The RNN-LSTM model had the strongest performance of the 3 models here as well. It achieved the lowest score for MAE as a % of the mean of x25. It's score was 11.97%. It also achieved the best RMSE as a % of the mean of x25, with a score of 19.94%. The RMSE score here for RNN-LSTM was noticeably lower than the Feed Forward Neural Network score of 22.57%, and significantly lower than the Linear Regression score of 24.26%. This proved that it had the smallest errors relative to the average value for x25. When looking at the MAE and RMSE as a % of the range of x25, the RNN-LSTM outperforms the other models here as well, obtaining a score of 2.59 for MAE as a % of the range of x25, and 4.15 for RMSE as a % of the range of x25. This proved that the RNN-LSTM model also had the smallest errors relative to the spread of x25 in the dataset.

6.1 FFNN Results

The evaluation also looked at the graphical fit of the predicted values against the true values. The graph in Figure 11 below shows the fitting of the predicted values from the Feed Forward Neural Network to the actual values of the x25 variable.

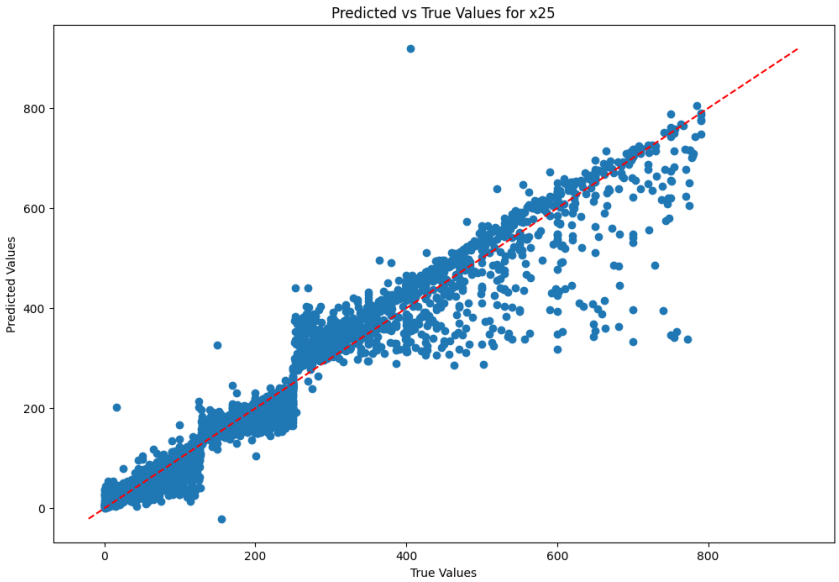

Figure 11: Feed Forward Neural Network Predictions

6.2 RNN LSTM Results

When evaluating the RNN LSTM results, it was important to monitor the loss and to ensure that the training loss does not deviate much from the validation loss. Figure 12 below shows the graph of training loss and validation loss during the fitting of the model. Due to the use of early stopping, the model stopped before 100 passes through the dataset.

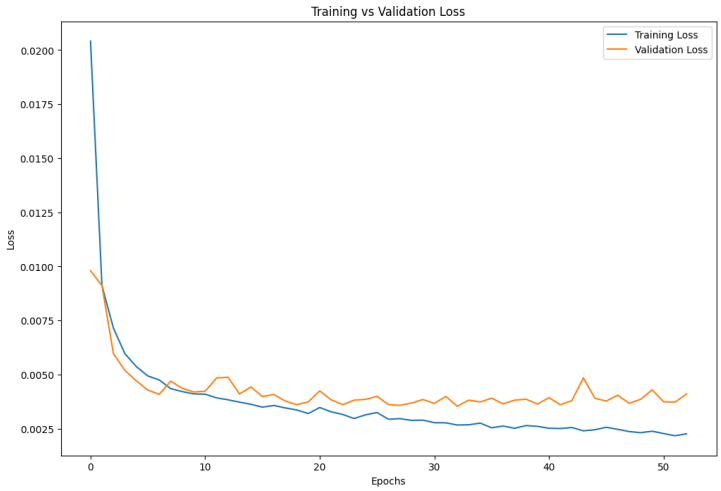

Figure 12: LSTM Training vs Validation Loss

Figure 13 below shows how the predicted values performed against the true values for the target variable. The values were inversely transformed after the normalization process to show the actual values.

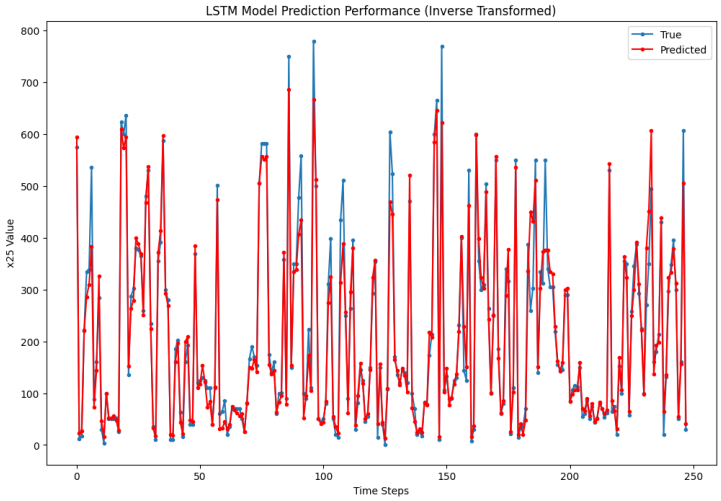

Figure 13: LSTM Predictions

6.3 LR Results

The fitting of the line for the Linear Regression model in Figure 14 below show that it is under-performing slightly when compared to the fitting for the Feed-Forward Neural

Network.

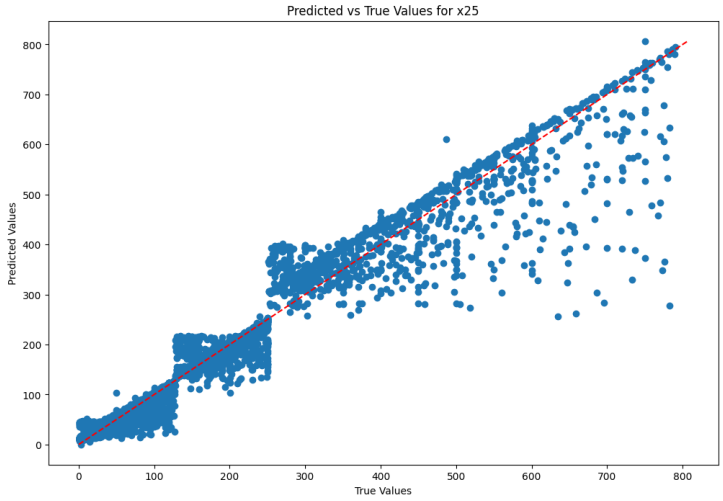

Figure 14: Linear Regression Predictions

6.4 Sub Questions Results

When viewing the correlation heatmap for the variables in the dataset, there was a very small correlation (45%) found between R&D spend ('x25') and Total Exports Sales ('x3). There was a more noticeable correlation (72%) between Total Export Sales and Total Material Costs ('x11'). There was also a correlation (62%) between Total Export Costs and Total Service Costs ('x14'). This suggests that there is a relationship between between the Export sales and the spend on Materials, Services and R&D. As Export Sales go up, there is an increased spend on R&D.
There was very little of a relationship identified between the Total Energy Cost ('x16a') and the spend on R&D. The correlation score was just 6%, suggesting that an increase in the cost of energy was not having an impact on the R&D spend.

7 Conclusion

The results of this project point towards a number of strong models for predicting the level of investment needed for R&D in Irish businesses. The trade-off for running a more intensive RNN LSTM model is that it achieves more accurate results. However, the Linear Regression model does get good results for a relatively simplistic model. The impact of these results will have a strong economic benefit for the companies involved, and can be used to better understand future results in the ABSEI survey. Companies will also be able to better plan their R&D investment strategy using the tools created in this report.
There was a large amount of work carried out to transform the data and reduce the dimensionality. The increased time needed to perform this work had a large positive impact on the results of the model. The work also allowed for better serving of the time series elements of the dataset and will benefit future work in this area.
Due to the exploratory nature of this project, there was a strong need to make sense of the raw data. The dataset brought challenges due to it's relative obscurity in the

academic field. The work done to transform and prepare the data will bring academic benefits for future projects by addressing core issues affecting it from a R&D perspective that can be used to analyse other parts of the survey.

7.1 Future Work

Future work in this area could look towards the different types of R&D, and the potential for that data to give greater accuracy for this model. Also, there is potential to look at other models to fit this data. From a time series perspective, an AutoRegressive Integrated Moving Average (ARIMA) model could be built to determine if it achieves better results than the RNN LSTM model. This was not chosen for this project due to the in its inability to make use of the other features of the dataset, and after observing work done by Hrnjica & Bonacci (2019) on the subject, which showed it to not be as effective as LSTM and Feed Forward Neural Networks.
Also, with greater knowledge now known about this dataset, there are a lot of other avenues for exploration. As stated in the sub research questions, there is a lot of work that could be done on the changing markets for exports, as well as the impact of changing energy costs and the general impact of climate change.

Acknowledgements

I would like to thank my supervisor, Jorge Basilio for his continued support and guidance throughout the process. I would also like to thank the Department of Enterprise, Trade & Employment for facilitating access to the data. Finally, I would like to thank my wife and kids for their patience and ongoing support, it would not have been possible to complete this without them.

References

Angenent, M. N., Barata, A. P. & Takes, F. W. (2020), Large-scale machine learning for business sector prediction, ACM, pp. 1143–1146.

Bansal, S. & Baliyan, N. (2022), 'Remembering past and predicting future: a hybrid recurrent neural network based recommender system', *Journal of Ambient Intelligence and Humanized Computing* .

Belitz, H. (2022), 'Research and development in german industry: High intensity, low growth'.
URL: *https://doi.org/10.18723/diw$_d$wr* : $2022 - 51 - 1$

Cao, J., Li, Z. & Li, J. (2019), 'Financial time series forecasting model based on ceemdan and lstm', *Physica A: Statistical Mechanics and its Applications* **519**, 127–139.

Fawcett, T. & Hampton, S. (2020), 'Why how energy efficiency policy should address smes', *Energy Policy* **140**, 111337.

Forino, G. & von Meding, J. (2021), 'Climate change adaptation across businesses in australia: interpretations, implementations, and interactions', *Environment, Development and Sustainability* **23**, 18540–18555.

gang Song, Y., lin Cao, Q. & Zhang, C. (2018), 'Towards a new approach to predict business performance using machine learning', *Cognitive Systems Research* **52**, 1004–1012.

Holl, A. (2021), 'The regional environment and firms' commitment to innovation: empirical evidence from spain', *Economics of Innovation and New Technology* **30**, 565–584.

Hrnjica, B. & Bonacci, O. (2019), 'Lake level prediction using feed forward and recurrent neural networks', *Water Resources Management* **33**, 2471–2484.

Jamshed, A., Mallick, B. & Kumar, P. (2020), 'Deep learning-based sequential pattern mining for progressive database', *Soft Computing* **24**, 17233–17246.

Kannan, R., Wang, I. Z. W., Ong, H. B., Ramakrishnan, K. & Alamsyah, A. (2021), 'Covid-19 impact: Customised economic stimulus package recommender system using machine learning techniques', *F1000Research* **10**, 932.

Kapoor, R. & Wilde, D. (2023), 'Peering into a crystal ball: Forecasting behavior and industry foresight', *Strategic Management Journal* **44**, 704–736.

Lee, G., Kim, D. & Lee, C. (2020), 'A sequential pattern mining approach to identifying potential areas for business diversification', *Asian Journal of Technology Innovation* **28**, 21–41.

Lee, J., Kwon, H.-B. & Pati, N. (2019), 'Exploring the relative impact of ramp;d and operational efficiency on performance: A sequential regression-neural network approach', *Expert Systems with Applications* **137**, 420–431.

Li, Z. (2023), 'Neural network economic forecast method based on genetic algorithm', *IET Software* .

Liu, E., Zhu, H., Liu, Q. & Udimal, T. B. (2022), 'Regional economic forecasting method based on recurrent neural network', *Mathematical Problems in Engineering* **2022**, 1–6.

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S., Reuter, U., Gama, J. & Gandomi, A. (2020), 'Data science in economics: Comprehensive review of advanced machine learning and deep learning methods', *Mathematics* **8**, 1799.

OTrakoun, J. (2022), 'Business forecasting during the pandemic', *Business Economics* **57**, 95–110.

Park, G. & Song, M. (2020), 'Predicting performances in business processes using deep neural networks', *Decision Support Systems* **129**, 113191.

Pinkse, J. & Gasbarro, F. (2019), 'Managing physical impacts of climate change: An attentional perspective on corporate adaptation', *Business Society* **58**, 333–368.

Reis, C., Ruivo, P., Oliveira, T. & Faroleiro, P. (2020), 'Assessing the drivers of machine learning business value', *Journal of Business Research* **117**, 232–243.

Stichhauerova, E., Zizka, M. & Pelloneova, N. (2020), 'Comparison of the significance of clusters for increasing business performance', *Journal of Competitiveness* **12**, 172–189.

Zhang, J. (2022), 'A neural network model for business performance management based on random matrix theory', *Mathematical Problems in Engineering* **2022**, 1–11.