BC2407 ANALYTICS II VISUAL & PREDICTIVE TECHNIQUES

Revitalizing LinkedIn:

A Machine Learning Approach to Enhance User Experience and Engagement

AY22/23 Sem 1 | Seminar 6, Team 5

NAME	MATRICULATION NUMBER
Jiang Lei	U2121557B
Toh Jing Qiang	U2121442H
David Tey Bo Yuan	U2121810J
Nagammai Senthil Kumar	U2120146L
Fong Ye Xuan	U2120767B

Click here to enter text.

Table of Contents

Click here to enter text.

Click here to enter text.

Click here to enter text.

Executive Summary

In the current dynamic job market, both job seekers and employers face significant challenges, including a widening gap between available talent and demand, uncertain future industry and skills demand, as well as the rise in job scams. LinkedIn is not spared from these challenges too. Fortunately, LinkedIn can capitalize on its unique position and vast repository of user data by implementing **IntelliLink**, a unified analytical solution designed to enhance security, promote innovation, and boost effectiveness in LinkedIn's job-seeking and recruiting ecosystem.

IntelliLink is a suite of machine learning models that will transform LinkedIn's platform by enhancing its **Talent Solutions**, **Premium Subscriptions**, and **LinkedIn Learning segments**. This is done via integrating three key features of IntelliLink:

(1) Unified Security through Job Scam Detection

With IntelliLink's fraudulent job listing Multinomial Naïve Bayes model that has high accuracy of 90% and low false negative rate of 26%, IntelliLink will enhance security on LinkedIn by automatically identifying and flagging fraudulent job postings for employees to verify. Additionally, a flagged fraudulent job listing will have key fraudulent words highlight to aid the employee in understanding why the job listing was flagged. Furthermore, IntelliLink's fraudulent job listing classification model provide insights as to which industries are common targeted and what commonly used strategies employed by scammers. For instance, fraudulent job listings are often targeted at inexperienced employees/students who are just starting out in their careers. This is especially so in low-barrier industries and functions like administration. With IntelliLink, LinkedIn's trust and credibility among its jobseekers and recruiters will be maintained, ensuring its legitimacy in the job market. Consequently, businesses and recruiters are more likely to use LinkedIn's Talent Solutions platform to find and hire talent, leading to increased profits for LinkedIn.

(2) Innovating with Industry Demand Forecasting

LinkedIn can distinguish itself from its competitors by harnessing the power of industry and skills demand forecasting models, specifically Taylor Expansion, which provides a high Mean Direction Accuracy score and interpretability. By implementing this model, LinkedIn can offer tailor-made courses for highly in-demand skills in the next 2 years, granting users and recruiters valuable insights that enable them to stay ahead of the curve and outpace their competition. With this innovation, LinkedIn can effectively unlock untapped revenue streams and maximize profitability at a relatively low cost. Overall, this forward-thinking approach cements LinkedIn's status as a true leader in the professional networking and career development space.

(3) Enhancing Effectiveness through Passive Jobseeker Identification

In order to unlock a vast pool of untapped talent for recruiters, LinkedIn can integrate a powerful head-hunter feature that enables recruiters to identify passive job seekers who may not be actively seeking employment but are open to new opportunities. Leveraging the capabilities provided by a random forest classifier that utilizes just 8 features and boasts a remarkable 91% accuracy and 83% recall score, LinkedIn can revolutionize the way recruiters approach talent acquisition. With an estimated 70% of the workforce comprising passive job seekers who may be invisible to traditional recruitment methods, this innovative feature positions LinkedIn as the ultimate destination for recruiters seeking to optimize their talent search process. By offering unparalleled access to this valuable segment of the labour market, LinkedIn can cement its position as the go-to platform for businesses looking to attract top-tier talent.

By addressing LinkedIn's existing shortcomings, IntelliLink improves the job-seeking and recruiting experience on LinkedIn, drive business profits, boost efficiency, and strengthen platform security. Therefore, empowering LinkedIn to remain competitive in the current dynamic job market.

1. Introduction

1.1 Current Situation of the Labour Market

In the current dynamic job market characterized by volatility, uncertainty, complexity, and ambiguity (VUCA), both employees and employers face daunting challenges. The statistics are staggering, with over 102,000 employees laid off from U.S.-based tech companies in 2023 alone (Ruby, 2023). Since the job market is highly dependent on industry shifts, global contexts, and national economics, job seekers face an uphill task of preparing for career changes. To stay competitive, companies must continually adjust their talent requirements, while job seekers must demonstrate adaptability and consistently update their skill sets to remain relevant amidst this constantly changing landscape (HRKatha, 2022). However, with constant changes in the job requirements, a widening gap has emerged between the available talents and job demand. Additionally, the increase in job vacancies after job layoff has further led to a rise in job scams, making it more challenging for job seekers to find employment.

Gap Between Available Talents and Demand

In Singapore 2022, the unemployment rate hovers around 3%, while the number of job vacancies is alarmingly high at 104,500 (Cue, 2023). These figures point towards a widening gap between the available talent and the demand for it in this ever-changing job market. As a result, job seekers are struggling to match their skills with changing job requirements, while organizations are finding it difficult to find the right talent for their positions.

Job Scams: A Growing Concern

With the high layoff rate, many people have turned to online job searching, leaving them vulnerable to job scams (Graham, 2020). In 2022, over 3,500 cases of job scams were reported, resulting in losses of over $58 million (Chua, 2022). This not only takes a toll on the job seekers' finances but also their mental well-being. Meanwhile, recruiters must spend more time building trust with job seekers and investing more money and effort to protect their company's reputation from these scams.

1.2 LinkedIn’s Position in Current Market

With the recent economic downturn and layoffs, the current job market serves as a golden opportunity for LinkedIn. With more people out of a job and companies needing to reduce costs and improve operational efficiency, there is a greater need to refine LinkedIn’s job matching services. LinkedIn can therefore take full advantage of this opportunity by addressing the biggest problems faced today by employers and job seekers. By doing so, it can solidify its position as the industry leader and drive innovation and growth in the job recruitment market.

While the current job market is advantageous to LinkedIn’s position, LinkedIn must simultaneously address the issue of job scams. Given that LinkedIn is one of the biggest job matching sites, it is a prime platform for scammers to target. LinkedIn has enjoyed its reputable status since it was founded and forged its brand as a legitimate job matching website, making it one of their most important assets. Therefore, LinkedIn must ensure that their platform continues to stay secure against these scammers and leverage technology to maintain their reputation and brand.

1.3 Opportunity Statement

LinkedIn can enhance its competitiveness and adaptability to industry trends by improving its framework of security, innovation, and effectiveness. By utilising its unique position, LinkedIn can leverage its vast repository of user data by implementing data analytics solution. With over 900 million registered members in more than 200 countries worldwide, LinkedIn is uniquely positioned to utilize machine learning algorithms to enhance user experiences and offer unparalleled value to both job seekers and recruiters (LinkedIn, n.d.).

2. IntelliLink – Unified Analytical Solution for Security, Innovation, and Effectiveness

IntelliLink is a comprehensive solution designed to enhance security, promote innovation, and boost effectiveness in LinkedIn's job-seeking and recruiting ecosystem. By integrating machine learning models to detect fraudulent job listings, identify passive jobseekers, and forecast industry trends, IntelliLink aims to address existing LinkedIn's shortcomings, upgrade LinkedIn's functionalities, and reinforce LinkedIn's competitive edge by utilizing its unique market position. IntelliLink achieves this by integrating the following 3 key features into LinkedIn.

(a) Unified Security through Job Scam Detection

LinkedIn's current method of combatting fraudulent job listings involves flagging suspicious accounts. While this is a useful approach, it does not directly address the issue of illegitimate job postings (Zamost & Khorram, 2022). It is entirely possible for a fraudulent job listing to be posted by a legitimate account, or for a scammer to create multiple fake accounts to post their scam job listings.

Therefore, even if suspicious accounts are flagged and removed from the platform, fraudulent job listings could still slip through the cracks. This means that LinkedIn's current system doesn't necessarily guarantee the legitimacy of job postings themselves, which could be a major concern for both job seekers and employers using the platform.

IntelliLink is therefore introduced to strengthen security on LinkedIn by utilizing natural language processing and classification models to directly flag suspicious job postings, rather than merely focusing on suspicious accounts. This approach fosters a secure and trustworthy environment, thus instilling confidence and building long-lasting relationships with LinkedIn's users.

(b) Innovating with Industry Demand Forecasting

LinkedIn currently encourages the sharing of industry news among the users so that the users can stay up to date about the latest news and developments in their field. However, industry news may not address individual needs as it covers broader trends and developments which may not always be applicable to individual users. Since LinkedIn does not offer any meaningful metrics on the market trends, users have to navigate the changing job market on their own (Rella, 2022).

IntelliLink takes LinkedIn's industry news sharing a step further by offering meaningful insights into the rapidly changing job market. This is achieved via IntelliLink's time series forecasting which analyzes and forecasts future industry outlooks and skill demand. Thus, job seekers are empowered to prepare better for future roles while enabling recruiters to source top-tier talent, distinguishing LinkedIn from its competitors.

(c) Enhancing Effectiveness through Passive Job seeker Identification

IntelliLink bridges the gap between recruiter demand and the 70% of the workforce made up of passive talents who have the necessary skills and qualifications to fill vacant positions but are not so visible to recruiters due to them not using LinkedIn's "Open to Work" feature on LinkedIn (Dewar, 2013).

Using machine learning models, IntelliLink predicts the interests of passive job seekers and unveils a vast, untapped talent pool. This innovative feature allows recruiters to allocate resources more efficiently and connect with candidates who may be open to opportunities but haven't explicitly indicated so on the platform. With IntelliLink, recruiters can enjoy higher successful placement rates thus establishing LinkedIn as the go-to platform for successful recruitment.

3. The Technology Behind IntelliLink

3.1 Fraudulent Job Listing Prediction

3.1.1 Methodology

To identify fraudulent job listings from legitimate ones, a credible dataset with 17879 rows (job listings) and 17 columns (features) from Kaggle was used. The modelling methodology is shown in *Fig. 3.1a*.

The '*fraudulent'* feature in the dataset is the target variable to be predicted.

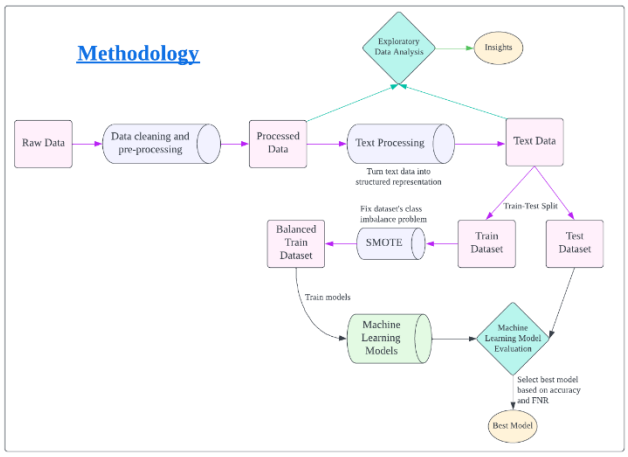


Fig. 3.1a.: Methodology used for building machine learning models.

3.1.2 Data Cleaning and Pre-processing

The dataset is cleaned by replacing N.A. values with empty strings instead of dropping rows with N.A. values as N.A. values does not represent any word in the context of Natural Language Processing. Dropping rows with N.A. values would not be a clear representation of all job listings as the missing values might be due to fraudulent intentions by the job listers, or due to the laziness of the job lister when posting many online job listings.

Since most of the variables of the raw dataset contains text variables and categorical variables with thousands of levels, the dataset is processed to generate a "text" column that is a combination of most of the other variables. Redundant columns are dropped as our analytical model will primarily focus on "text" data since job listings are mainly text-based. A snippet of this processed dataset is shown in *Fig. 3.1b*.

	telecommuting	has_company_logo	has_questions	fraudulent	text
0	0	1	0	0	Marketing Intern US, NY, New York Marketing ...
1	0	1	0	0	Customer Service - Cloud Video Production NZ,...
2	0	1	0	0	Commissioning Machinery Assistant (CMA) US, I...
3	0	1	0	0	Account Executive - Washington DC US, DC, Was...
4	0	1	1	0	Bill Review Manager US, FL, Fort Worth Spot...

Fig. 3.1b: Snippet of processed dataset for training analytical models

3.1.3 Text Processing

The "text" data by itself is unstructured and cannot be passed into most analytical models. Thus, there is a need to convert text into numerical representation via text vectorization. This makes the "text" data structured.

Additionally, raw "text" data contain useless stop words like "a", and "is", and contain words that have numerous representations (e.g.: "apple" and "apples" have the same root word but are viewed as different words by analytical models). Hence, the "text" data need to be processed before using it to train analytical models. The text processing pipeline is shown in *Fig. 3.1c*.

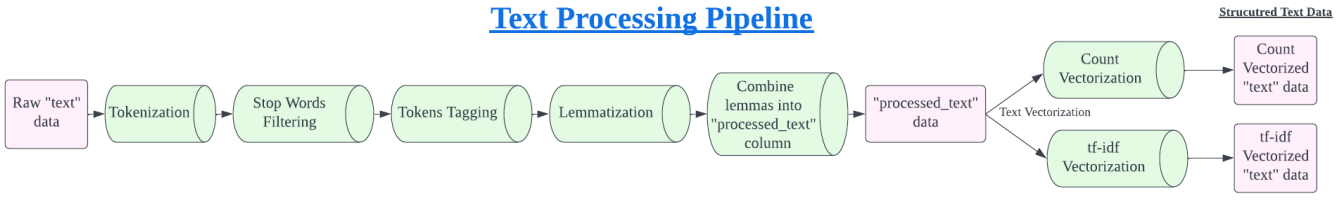


Fig. 3.1c: Text processing pipeline

Click here to enter text.

3.1.4 Exploratory Data Analysis

By visualising the frequent words that appeared in both fraudulent and non-fraudulent job listings respectively *(Fig. 3.1d)*, it was observed that fraudulent job listings contain words like “Entry level”, “data entry”, “high school”, “support”, “project”, and “design” frequently, while non-fraudulent job listings contain words like “Full time”, “bachelor’s degree”, “Information Technology”, and “Senior” frequently. This pattern suggests that fraudulent job listings are often targeted at inexperienced employees/students who are just starting out in their careers. This pattern is strongly supported by exploratory data analysis on individual features *(Appendix A1)*.

Fig. 3.1d: Word Clouds of non-fraudulent job listings (left) and fraudulent job listings (right)

Additionally, fraudulent job listings tend to have lower character and word counts *(Fig. 3.1e)*.

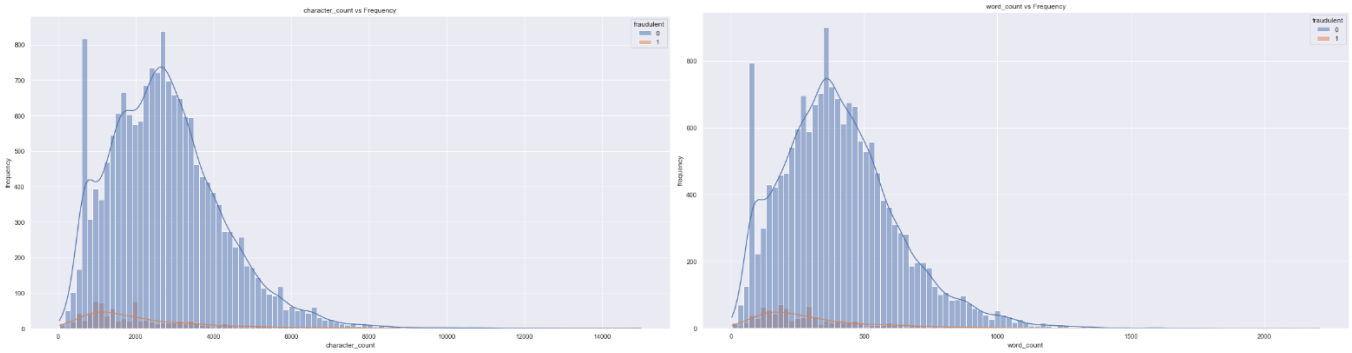


Fig. 3.1e: Histogram Plots of Character Counts (left) and Word Counts (right) for fraudulent job listings (red) and non-fraudulent job listings (blue)

3.1.5 Model Training and Evaluation

(a) Fixing dataset class imbalance problem

Since the dataset is extremely unbalanced, with only 4.84% of the rows being labelled as fraudulent and the remaining 95.16% labelled as non-fraudulent *(Fig. 3.1f)*. Thus, the train dataset was resampled using **Synthetic Minority Oversampling Technique** (SMOTE) to ensure an even proportion of fraudulent and non-fraudulent cases in the train dataset. This is to incentivise the model to learn how to predict fraudulent cases accurately. Hence, reducing false negative rate and increasing prediction accuracy for fraudulent cases.

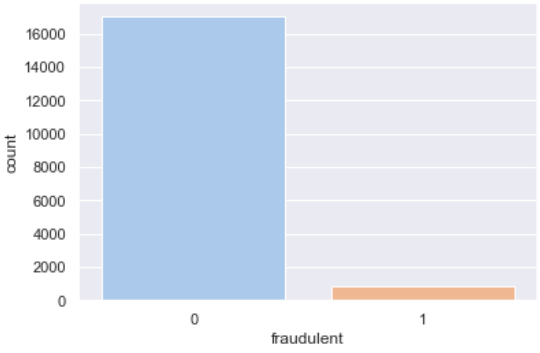


Fig. 3.1f: Distribution of “fraudulent” in dataset

Click here to enter text.

(b) Defined Metrics to Evaluate Model Performance

Classification accuracy and false negative rates are the key metrics for evaluating model performance for fraudulent job listings classification. **High classification accuracy** of **over 85%** to ensure that predictions are accurate in general and instil trust in the model by users, and **low false negative rate** of **less than 30%** to ensure that fraudulent job listings do not get away undetected.

(c) Training and Testing Machine Learning Models

The processed and vectorized "text" data is used to train and test 4 different classification models – **Multinomial Naïve Bayes Classifier**, **Logistic Regression**, **Support Vector Classifier** and **Random Forest**. Each model is trained on both Count Vectorized and TF-IDF Vectorized "text" datasets, tested using the test dataset and evaluated using **classification accuracy** and **false negative rate**.

The model test performance results for each model are shown in *Fig. 3.1g*. More details about the individual model performances can be found in *Appendix A2*.

Model	Classification Accuracy	False Negative Rate
Multinomial Naive Bayes Classifier (TV)	0.8999	0.2643
Multinomial Naive Bayes Classifier (CV)	0.8958	0.2731
Logistic Regression (TV)	0.9726	0.5507
Logistic Regression (CV)	0.4614	0.0396
SVC (TV)	0.8850	0.6916
SVC (CV)	0.6152	0.4626
Random Forest (CV)	0.2604	0.0088

Fig. 3.1g: Model test performance results for different models

3.1.6 Model Selection – Multinomial Naïve Bayes Classifier with td-idf Vectorization

Based on the model performances on the same test dataset, the Multinomial Naïve Bayes Classifier with td-idf vectorization performs the best as it has a very high classification accuracy of about 90% and relatively low false negative rate about 26%. Therefore, this model is used to predict potential fraudulent job listings on LinkedIn.

3.1.7 Feature Importance – Fraudulent Words

Using the selected Multinomial Naïve Bayes Classifier model, the important features, i.e.: words that contribute the most to a job listing being fraudulent, is obtained from the model *(Fig. 3.1h)*. The important fraudulent words include "jacksonville job descriptionadminstrative", "restaurant manager awarded", "office manager pi", "portland sales need", etc. These fraudulent words allow LinkedIn employees to understand more clearly what causes a job listing to be classified as fraudulent.

words	real	fraudulent	fraudulent_ratio
jacksonville job descriptionadministrative	0.000084	0.001255	14.908364
restaurant manager awarded	0.000084	0.001202	14.270381
office manager pl	0.000084	0.001162	13.801295
portland sales need	0.000084	0.001134	13.465405
philadelphia administrative 21	0.000084	0.001058	12.570614

Fig. 3.1h: Words that contribute most to a fraudulent job listing.

3.1.8 Easy Model Integration with LinkedIn's Backend Services

The model and vectorizer files (.sav extension) are available for download and import into LinkedIn's backend services *(Fig. 3.1i)*. This allows the model to be integrated easily onto LinkedIn's platform without writing much code. The input text data can consist of anything related to the job listing. For example, job title, location, department, salary range, company profile, description, requirements, benefits, employment type, required experience, required education, industry, function, etc. All these features of the job listing will be used to predict and determine if the job listing is fraudulent.

```
# Import
import joblib
from models.execute_tv_model import execute_tv_model

# load the model from disk
loaded_mnb_tv_model = joblib.load('models/mnb_tv_model.sav')
tfidf_vectorizer = joblib.load('models/tf-idf_vectorizer.sav')

# Execute the model
prediction_random_text = execute_tv_model(text_data_random, loaded_mnb_tv_model, tfidf_vectorizer)

# Print Results
print("Input text data:", text_data_random)
print("Prediction:", prediction_random_text)
```

Fig. 3.1i: Steps to import the model and try it out yourself.

Click here to enter text.

3.2 Industry Demand Forecasting

3.2.1 Methodology

To forecast the job industry demand and skills demand of the future, we will utilize time series forecasting techniques on sample datasets from SingStat and WorldBank. Our approach will begin with the Autoregressive Integrated Moving Average (ARIMA) model and evaluate its suitability for LinkedIn's context. We will also explore the Holt-Winters Model and Taylor Expansion Model, assessing their accuracy and robustness. Our forecasting horizon will be 8 quarters, which strikes a balance between preparing job seekers and model precision.

3.2.2 Dataset

(a) Job Vacancy by Industry

Our analysis will leverage Singapore's quarterly job vacancy data, sourced from SingStat. The dataset features a range of industries in Singapore, segmented into various levels, with level 3 offering the most specific classification. In total, we will evaluate 32 industries to predict their future demand.

(b) Skills Demand by Industry

Additionally, we will incorporate a dataset from WorldBank that highlights the most in-demand skills for each industry per year. This dataset is generated via web-scraping LinkedIn's job postings, followed by text mining to track the frequency of specific keywords and skills. The resulting information is used to rank skill importance. This dataset will be merged with forecasting analysis to anticipate which skills will be in high demand across all industries.

3.2.3 ARIMA Model

The first model we used is the ARIMA model. This model assumes that future data points are a linear combination of previous data points and previous noise variables. The time series may also be differenced, for instance, to predict the rate of change of the time series. This may help improve the model as it may be easier to predict the rate of change compared to the original values.

The Auto Arima algorithm is used to choose the best model parameters, namely the lookback period for the autoregressive part of the model, the number of times the time series is differenced, and the lookback period for the moving average part of the model. The Akaike Information Criterion (AIC) is used as a metric for model performance, and the best model is chosen based on this metric.

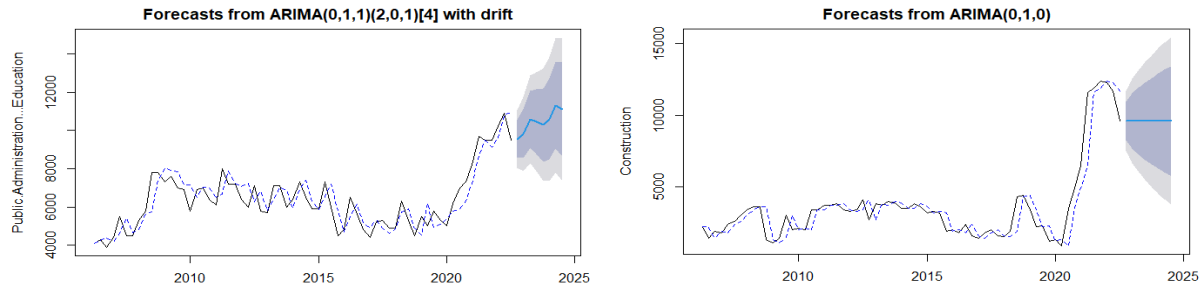


Fig. 3.2a Forecasts from ARIMA, 2 of 32 industries (Refer to Appendix B3.1 for all plots)

While ARIMA models may work exceptionally well for certain datasets, we can see that for many of our time series, the Auto Arima model chose the null model as the optimal model. This means that the best predictor of future value is simply the current value. One reason may be due to the regime change in the data. For example, exogenous events such as the COVID-19 pandemic heavily affected Singapore's economy, which is reflected in the data by the spike in job vacancies. Since ARIMA assumes that the underlying process that generates the data is constant through time, it is not able to adapt to sharp changes due to exogenous events.

3.2.4 Holt-Winters Model

To address this issue, we used the Holt-Winters model instead. The exponential smoothing used in the Holt-Winters model helps weigh the most recent data points more heavily, making it more suitable for modelling a dynamic system. The decomposition of the original time series into trend and seasonal components also makes it much more intuitive compared to ARIMA models.

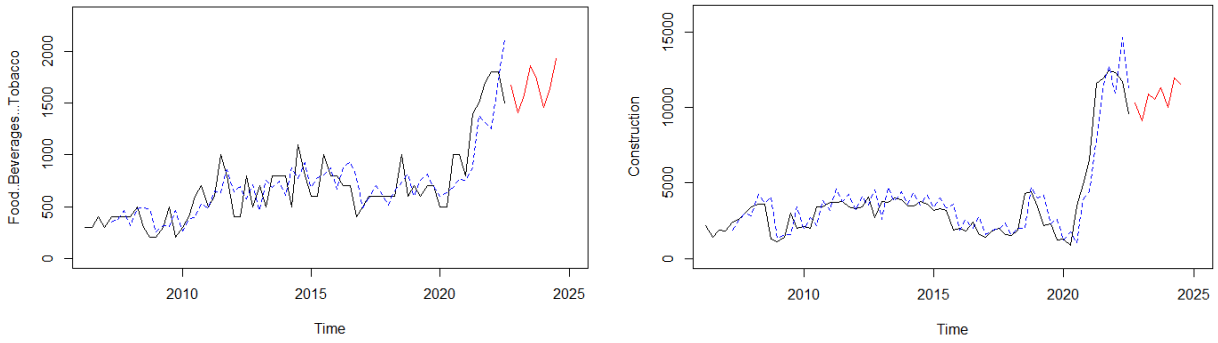


Fig. 3.2b Forecasts from Holt-Winters Model, 2 of 32 industries (Refer to Appendix B4.1 for all plots)

While the Holt-Winters model captures the trend decently well, the fixed seasonality results in the model overfitting the noise, resulting in seasonal predictions even when there are no clear seasonal trends in the original data.

3.2.5 Taylor Expansion Model

Therefore, we applied a new model that serves to provide an intuitive interpretation of its forecasts. Instead of decomposing the time series into its trend and seasonal components, we try to estimate the instantaneous trend and acceleration, and use this to extrapolate to future data points. We accomplish this by estimating the derivatives of the time series, and then calculating the future estimates based on Taylor Series Expansion.

The model takes in 5 parameters:

Parameters	**Description**
Lag	Number of data points to average in the estimation of derivatives
Degree	The degree of approximation used (number of derivatives). E.g. degree = 1 means the model uses linear approximation, degree = 2 means that the model uses quadratic approximation
Freq	Frequency of time series data
Forecast Period	Number of periods to forecast out of sample
Time Step	Amount of time between each data point, used to estimate the derivatives accurately

Fig. 3.2c: Table of model parameters and its descriptions

The 2 most important parameters are lag and degree. Since the lag parameter affects how far to look to estimate the current instantaneous derivatives, a smaller lag parameter would mean that the model is more sensitive to recent changes in trend. A longer lag parameter would decrease its sensitivity, therefore acting as a smoothing function.

The degree is the number of derivatives used to approximate the forecasts. The higher the degree, the more terms are estimated and used for forecasting. Therefore, we can model curvature in the data as well as some seasonal behaviours

in the short run. However, as we increase the degree, the model becomes more susceptible to input noise and may become unstable. For a more technical explanation, refer to Appendix B5.1.

In our case, we can use the first and second derivatives to estimate the trend and acceleration of the demand in different industries. Intuitively, our model assumes that the current trend and acceleration of industry demand will remain constant into the future and use these values to extrapolate future demand.

As the confidence interval of time series models generally increase significantly the further we forecast into the future, only nearby forecasts are reliable enough to be used in practice. Hence, the ability to extrapolate seasonal patterns far into the future makes little difference in practice. This is true especially for our data, whereby there are no clear seasonal trends, and the underlying driver of our data is constantly changing.

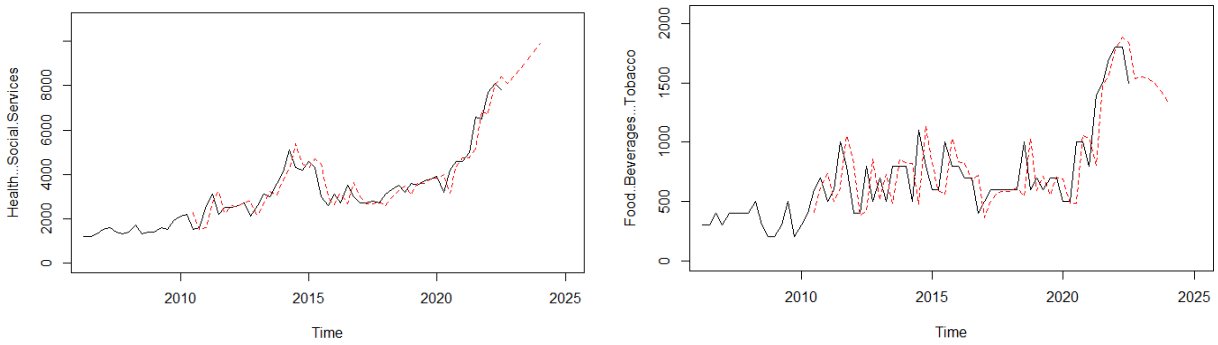


Fig. 3.2d Forecasts from Taylor Expansion Model, 2 of 32 industries (Refer to Appendix B5.3 for all plots)

We can see that this model is much more sensitive to changes in the underlying trend of the different industries' demand.

3.2.6 Model Evaluation

To evaluate the three time series models, we will use three different metrics, the root mean squared error, mean absolute percentage error and the mean direction accuracy.

Model	**RMSE**	**MAPE**	**MDA**
ARIMA	382.84	0.2107	0.6366
Taylor Expansion	452.39	0.2499	0.6341
Holt-Winters	1485.79	2.6021	0.4536

Fig. 3.2e: Table of model performance metrics

We can see that for all three metrics, the ARIMA was the best, followed closely by the Taylor Expansion model and finally the Holt-Winters model. While the Holt-Winters model is very intuitive to use, its performance is much worse compared to the other two models. Therefore, the Holt-Winters model should not be used.

The performance of the ARIMA model and the Taylor Expansion model is very similar, with the same mean direction accuracy up to 2 decimal places. However, there are some drawbacks of using ARIMA even though it performed better according to these metrics.

Firstly, its model assumes that the whole time series is generated by the same underlying process, which is unrealistic as the job market is constantly changing. We can clearly see the effect of COVID-19 on our data, which would result in the coefficients estimated by the ARIMA model irrelevant if it was fit on past data.

Secondly, it is difficult to understand the forecasting behaviour of the model by simply looking at the coefficients. Being a purely statistical model, ARIMA simply finds the coefficients that minimises its error with the true value. The Taylor Expansion model, by comparison, is extremely interpretable as the current trend and acceleration estimates can be easily displayed by the model. Due to the high error margins in time series forecasting, forecasts are typically treated as baseline values in practice. Therefore, the additional interpretability can help users combine the model's forecast with external factors to come up with their own best guess.

For example, if the cloud services industry experienced a boom in the last year, the model may extrapolate this and predict that next year would be even better. However, one may also notice contradictory evidence, perhaps that investment in new cloud services start-ups has been drying up due to a slowdown in technological breakthroughs. Thus, the user may combine his or her own domain knowledge together with the model's baseline prediction to make a more informed decision.

Therefore, the Taylor Expansion model is chosen as the benefits of increased interpretability and intuition outweighs the slightly reduced prediction performance compared to the ARIMA model.

3.2.7 Skills Forecasting

By mapping the top skills in demand in each industry to our forecast of future industry demand, we can calculate which skills are likely to be in demand in the future using matrix multiplication. This can help LinkedIn users better prepare themselves for the job market and priorities their time to develop the most valuable skills. For more details on how the skill demand is calculated, please refer to Appendix B6.

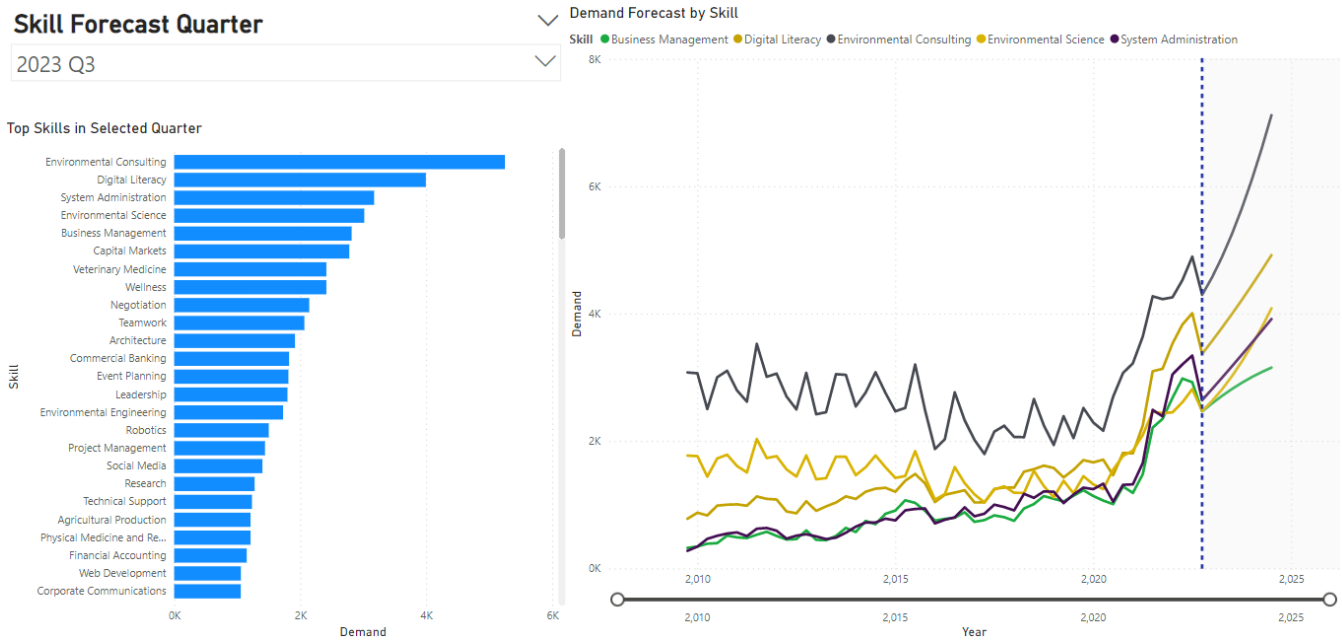


Fig 3.2f & 3.2g: Top skills in 2023 Q3 (left), Top 5 skills forecast over time (right)

For example, the top 5 skills forecasted in 2023 Q3 are Environmental Consulting, Digital Literacy, System Administrator, Environmental Science and Business Management. The trend can also be seen on the plot, which can be used to compare the demand of various skills in the future. Therefore, we can see that environmental and digital skills will become more important in the future, allowing LinkedIn users to prepare themselves by learning these relevant skills.

3.3 Job Seeker Prediction

3.3.1 Methodology

In order to distinguish a jobseeker from a large set of user profiles, a credible dataset was utilized from Kaggle comprising of 21287 rows and 15 columns. These columns contain several essential features such as the city development index, user training hours, and current employment situation, while some unrelated columns like index and enrollee_id were removed during the data cleaning process. The modelling pipeline utilized is illustrated in *Fig. 3.3a*. Notably, the dataset's 'target' feature represents the variable to be predicted, where a value of 1 indicates a jobseeker, whereas a value of 0 corresponds to a non-jobseeker.

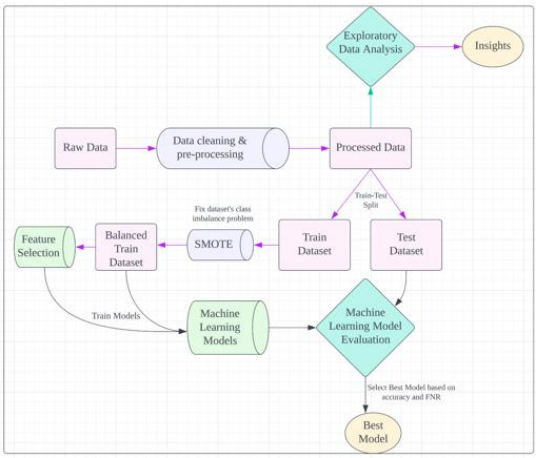


Fig. 3.3a.: Methodology used for building machine learning models.

3.3.2 Data Pre-processing & Exploration

(a) Deal with unbalance dataset

The dataset is initially imbalanced with a 1:3 proportion (Job seekers: Non Job seekers), which may lead the model to exhibit bias towards predicting the majority class, i.e., non-jobseekers. Hence, oversampling technic, SMOTE is employed to reduce false negative rates.

(b) Deal with missing data

To avoid making inaccurate assumptions during imputation, the 10% of missing values in the targeted predictor variable were dropped.

Additionally, several features, such as *company size*, *company size*, *education level*, *major discipline*, and *enrolled university*, were determined to be highly relevant to the target predictor of whether an individual is a job seeker and displayed a high correlation with other variables in the dataset. For example, individuals pursuing STEM majors were more likely to hold a graduate or master's degree. However, since around 30% of the dataset's rows lacked data in these columns, filtering them out would have resulted in significant data loss. Thus, missing values in these crucial variables were imputed using random forest. In the case of the gender variable, since the missing values might be due to privacy concerns from user profiles, they were imputed as a separate category - "*unknown*".

(c) Feature Engineering

As some user profile features, such as current employment information, may be missing due to privacy concerns and real-world impracticality, we narrowed down the available 18 features to the top 8 most important ones using four tree-based models *(Fig. 3.3b)*. Although Logistic Regression with Recursive Feature Elimination is commonly utilized to reduce feature dimension, it demonstrated limited explanatory power with a low R-square value of 0.56 on this dataset; thus, we disregarded its significant features.

	CART	ETC	RF	GBC	Mean_Importance
city_development_index	1.000000	1.000000	1.000000	1.000000	1.000000e+00
enrolled_university	0.074737	0.302406	0.320148	0.071649	1.922349e-01
experience	0.125456	0.228329	0.094200	0.021016	1.172502e-01
training_hours	0.180497	0.211963	0.005438	0.007894	1.014482e-01
company_size	0.073754	0.146569	0.024495	0.021588	6.660153e-02
last_new_job	0.050246	0.096085	0.007680	0.002006	3.900443e-02
education_level	0.036282	0.077004	0.020070	0.010175	3.588251e-02
company_type	0.033399	0.077780	0.022157	0.006213	3.488718e-02
relevent_experience	0.014830	0.036617	0.022113	0.001894	1.886334e-02
major_discipline_STEM	0.004435	0.030379	0.037933	0.002485	1.880774e-02

Fig. 3.3b: Top 10 important features identified in each model.

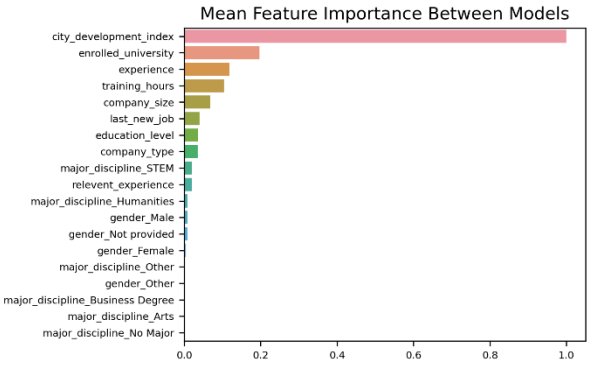


Fig. 3.3c: Rank of averaged feature importance

Click here to enter text.

The chosen features were determined to be the most significant in identifying passive job seekers, and they were consistent with the results from exploratory data analysis (EDA).

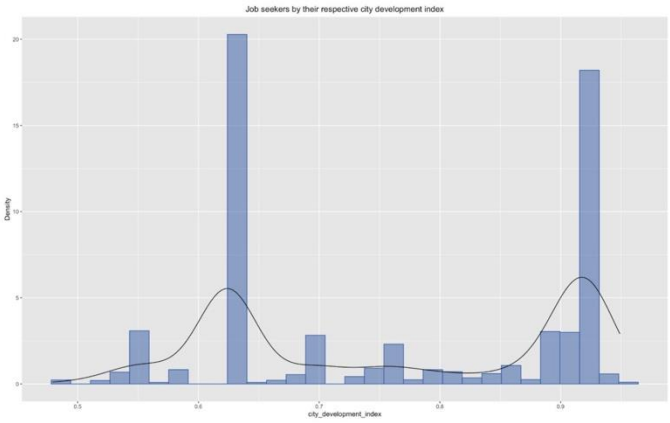


Fig. 3.3d: Jobseekers' City Development Index Distribution

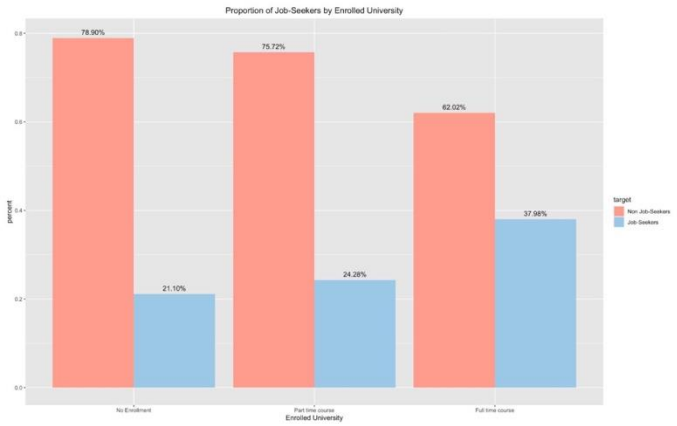


Fig. 3.3e: Proportion of jobseekers & Non jobseekers by education type

The city development index was found to be the primary driver of job-seeker prediction (*Fig. 3.3c* and EDA revealed that jobseekers were predominantly found in the least developed or well-developed cities *(Fig. 3.3.d)*, possibly due to limited job opportunities and career advancement in the former and a greater motivation to explore career changes in the latter. Additionally, the type of university attended was found to be second critical factor in predicting job changes *(Fig. 3.3.c)*, with full-time graduates being more likely to seek out new job opportunities, in line with EDA finding in *Fig 3.3.f*. This may be attributed to the significant investment of time and money that full-time graduates make in their education, leading to a greater sense of ambition and responsibility to maximize their degree.

Thus, the reduced dataset consisting of 8 most important features, will serve as a secondary dataset to test model's performance with a more limited number of variables, in addition to the initial dataset that includes all 18 features.

3.3.4 Model Development

(a) Defined Metric of Model Performance

Misclassifying jobseekers as non-jobseekers can result in missed opportunities, while misclassifying non-jobseekers as jobseekers can waste recruiters' time. To prevent these errors and maintain user trust, overall accuracy is set as the primary metric with a benchmark of 90%.

However, relying solely on accuracy is not sufficient. To safeguard potential revenue from successful placements and prevent the undermining of user trust in LinkedIn's matching capabilities, minimizing the error of misclassifying job seekers as non-job seekers is crucial. Hence, the recall score is defined as a secondary benchmark to prioritize a lower false-negative rate while maintaining overall performance.

(b) Training Process

The training process involved 5 models for classification: three from tree families - Classification and Regression Trees (CART), Random Forest, and Extreme Gradient Boosting (XGB), as well as Support Vector Machine (SVM) and logistic regression (LR). Each model was trained on 2 datasets with hyperparameter tuning and 5-fold cross validation:

1) the initial dataset containing 18 features with tuned parameters.
2) the selected dataset containing only 8 features with tuned parameters.

3.3.5 Model Selection

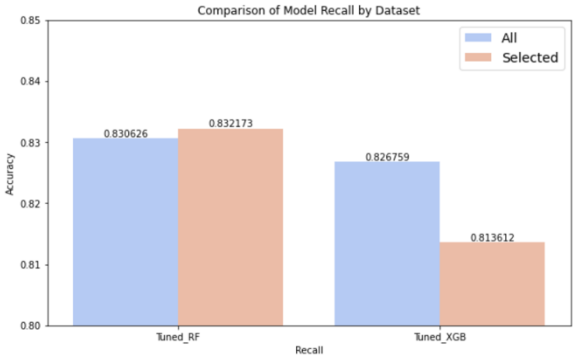
After the first benchmark (accuracy > 90%), it was observed that LR and SVM models had low accuracy scores *(Fig. 3.3.f)*, and hence were excluded from further competition. In contrast, the models belonging to the tree families exhibited comparable performance, with only a minor difference of 1% in their accuracy scores.

	Model	Dataset	Accuracy	TPR	TNR	FPR	FNR	Precision	Recall	F1 Score
10	Tuned_XGB	All	0.914354	0.826759	0.941430	0.058570	0.173241	0.813546	0.826759	0.820100
4	Tuned_RF	All	0.910884	0.830626	0.935692	0.064308	0.169374	0.799702	0.830626	0.814871
11	Tuned_XGB	Selected	0.910701	0.813612	0.940712	0.059288	0.186388	0.809231	0.813612	0.811415
5	Tuned_RF	Selected	0.908145	0.832173	0.931628	0.068372	0.167827	0.790015	0.832173	0.810546
1	Tuned_CART	All	0.900840	0.816705	0.926847	0.073153	0.183295	0.775330	0.816705	0.795480
2	Tuned_CART	Selected	0.893353	0.843774	0.908678	0.091322	0.156226	0.740665	0.843774	0.788865
13	Tuned_LR	All	0.865595	0.876257	0.862300	0.137700	0.123743	0.662961	0.876257	0.754830
14	Tuned_LR	Selected	0.848064	0.832173	0.852976	0.147024	0.167827	0.636310	0.832173	0.721180
7	Tuned_SVM	All	0.799489	0.788090	0.803012	0.196988	0.211910	0.552903	0.788090	0.649872
8	Tuned_SVM	Selected	0.796384	0.773395	0.803490	0.196510	0.226605	0.548847	0.773395	0.642055

Fig. 3.3.f: Ranking of model performance based on accuracy.

	Model	Dataset	Accuracy	TPR	TNR	FPR	FNR	Precision	Recall	F1 Score
5	Tuned_RF	Selected	0.908145	0.832173	0.931628	0.068372	0.167827	0.790015	0.832173	0.810546
4	Tuned_RF	All	0.910884	0.830626	0.935692	0.064308	0.169374	0.799702	0.830626	0.814871
10	Tuned_XGB	All	0.914354	0.826759	0.941430	0.058570	0.173241	0.813546	0.826759	0.820100
1	Tuned_CART	All	0.900840	0.816705	0.926847	0.073153	0.183295	0.775330	0.816705	0.795480
11	Tuned_XGB	Selected	0.910701	0.813612	0.940712	0.059288	0.186388	0.809231	0.813612	0.811415

Fig. 3.3.g: Ranking of models based on recall score with accuracy > 90%



Among the top two performers, it is observed that random forest models outperform XGB models with a higher recall score, and lower false negative rate *(Fig. 3.3.g)*. Moreover, random forest works better on both initial dataset with 18 features and filtered dataset with selected 8 features *(Fig. 3.3.h)*, indicating its robustness and suitability for the job-seeker prediction task, especially when having to deal with limited available features.

Fig. 3.3.h: Random Forest and XGB Recall Score Comparison on Various Datasets

3.3.6 Selected Model Evaluation

The optimal model for jobseeker classification is the random forest model, trained on the selected dataset of 8 features. It has an accuracy rate of 91% and the highest recall score of 0.83, making it a reliable tool for LinkedIn recruiters to predict passive job seekers consistently. The model's stability was verified through 10-fold cross-validation, which demonstrated low standard deviation across different folds *(Fig. 3.3i)*. Therefore, recruiters can count on the model's recommendation of job seekers who are most likely to accept their job offers.

Moreover, with an increase in the number of training examples, the cross-validation score shows a positive trend of improvement *(Fig. 3.3j)*, indicating that the model is expected to perform well on unseen data. Hence, when LinkedIn utilizes the model for predicting jobseekers, the outcomes can be utilized to improve the model in the future.

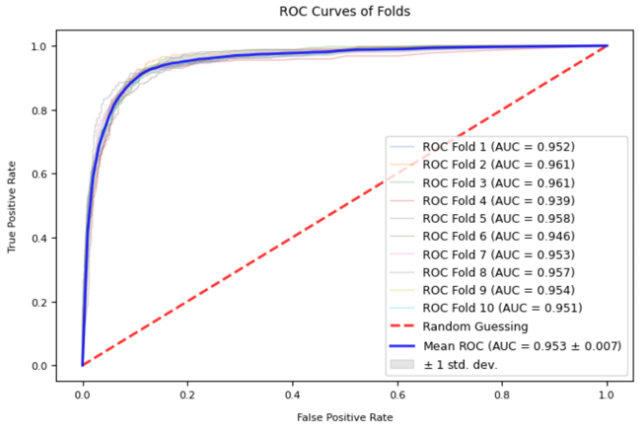


Fig. 3.3i: ROC Curves of Folds (10-fold cross validation)

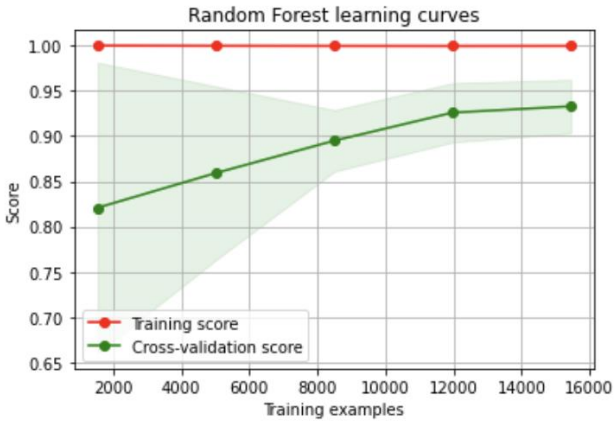


Fig. 3.3j: Cross validation score vs Training examples

We delved deeper into the model's prediction confidence *(Fig. 3.3k)* and observed that the correctly predicted jobseekers had a median prediction probability of 0.99 and an interquartile range (IQR) ranging from 0.94-0.95, whereas the IQR for incorrectly predicted jobseekers was 0.73-0.95. These two ranges were found to have little to

Click here to enter text.

no overlap, which can be utilized by LinkedIn to prioritize recommended passive jobseekers with a high probability of prediction (>0.95). The recommendation index based on probability can be made available to recruiters, enabling them to make more informed decisions.

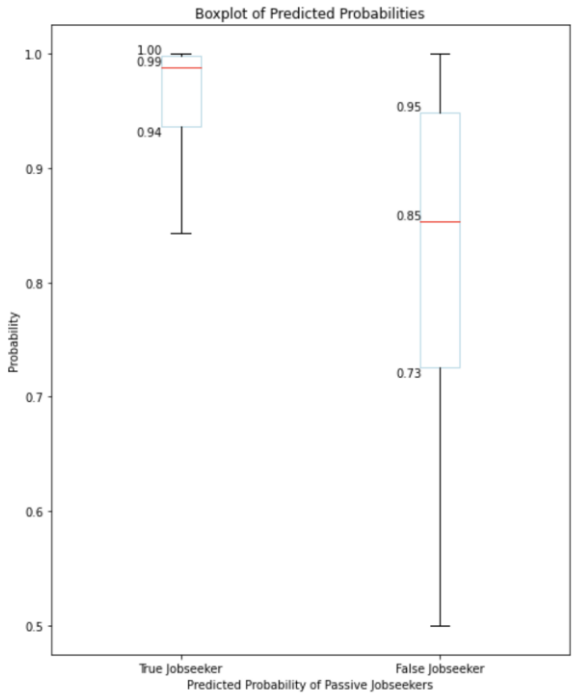


Fig. 3.3k: Boxplot of predicted probabilities of passive jobseekers

Click here to enter text.

4. Linking IntelliLink with LinkedIn

4.1 IntelliLink for LinkedIn Users

With IntelliLink's 3 different machine learning models for predicting fraudulent job listings, forecast industry trends, and identify passive job seekers, LinkedIn can make the recruitment experience much better for both job seekers and recruiters *(Fig. 4.1a)*.

For job seekers, the fraudulent job listing prediction ensures that the jobs that they have applied to are credible, hence they won't get scammed by malicious parties. Additionally, if the job seekers are not successful in landing a job from LinkedIn, they can learn the trending relevant skills as revealed by the industry trend forecast. This increases their chance of being hired now and in the future. Furthermore, job seekers are also part of the passive job seeker pool on LinkedIn which enables them to get headhunted by recruiters passively.

For recruiters, the fraudulent job listing prediction makes them more credible and trustworthy to the job seekers. This boosts the application rate for the recruiter. Additionally, since the job seekers have relevant skills from the industry trend forecasting, the quality of job seekers applying for their jobs will be higher. Thus, resulting in a higher chance of hiring a better candidate. Furthermore, recruiters can increase their applicant pool size and chance of offer acceptance by using IntelliLink's passive job seeker identification model that identifies passive job seekers that fits what the recruiters are looking for. This increases recruitment efficiency and reduce costs.

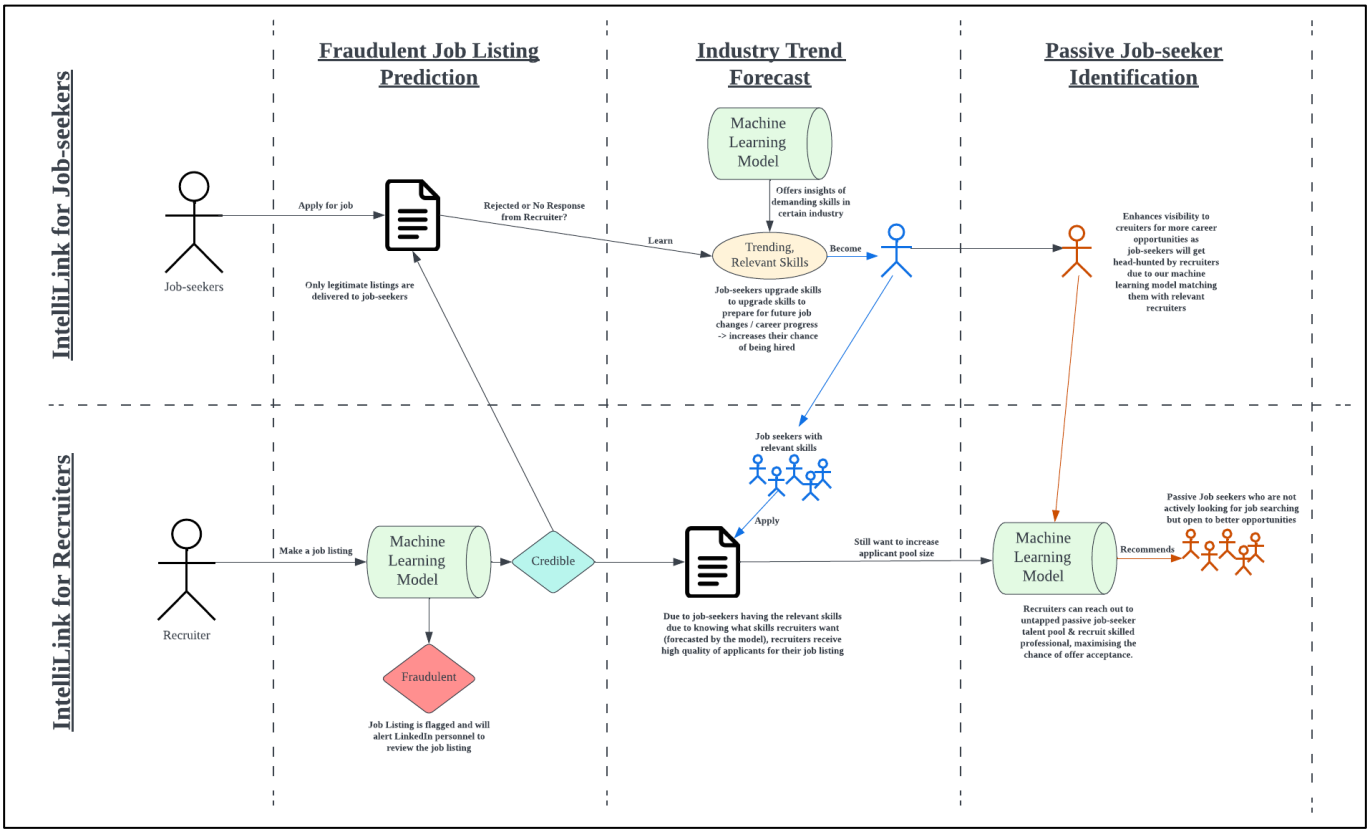


Fig. 4.1a: User Journey Diagram for job seekers and recruiters

Click here to enter text.

4.2 Integration with LinkedIn

IntelliLink can integrate well into LinkedIn's existing operations by improving operations in their Talents Solutions, Premium Subscriptions and LinkedIn Learning.

4.2.1 Fraudulent Job Listings Prediction Integration

Our fraudulent job listing prediction model can be integrated into LinkedIn's **Talent Solutions division** to (a) automatically flags out suspicious job postings, and (b) provides LinkedIn with insights on fraudulent job postings *(Fig. 4.2a)*.

(a) Automatically Flags out Suspicious Job Postings

IntelliLink's fraudulent job classification model identifies and flags suspicious new job postings for review by supervisors. If a job posting is deemed fraudulent, it will be removed, and the recruiter's account will receive a strike, with three strikes resulting in account suspension. The fraudulent postings also serve as data to help enhance model efficiency and accuracy in this dynamic job market where job scams are becoming more creative.

(b) Provides LinkedIn Insights on Fraudulent Job Postings

Flagged job postings display highlighted key words that are likely to contribute to the job posting being fraudulent. This helps LinkedIn employees understand why the job posting was flagged and make decision in determining whether the job listing is fraudulent. Additionally, LinkedIn employees can monitor trends and patterns in fraudulent postings as displayed on LinkedIn employee admin dashboard to identify vulnerable industries or parties, and develop targeted strategies, like stricter checks for the accounting industry, to combat scams and protect users effectively.

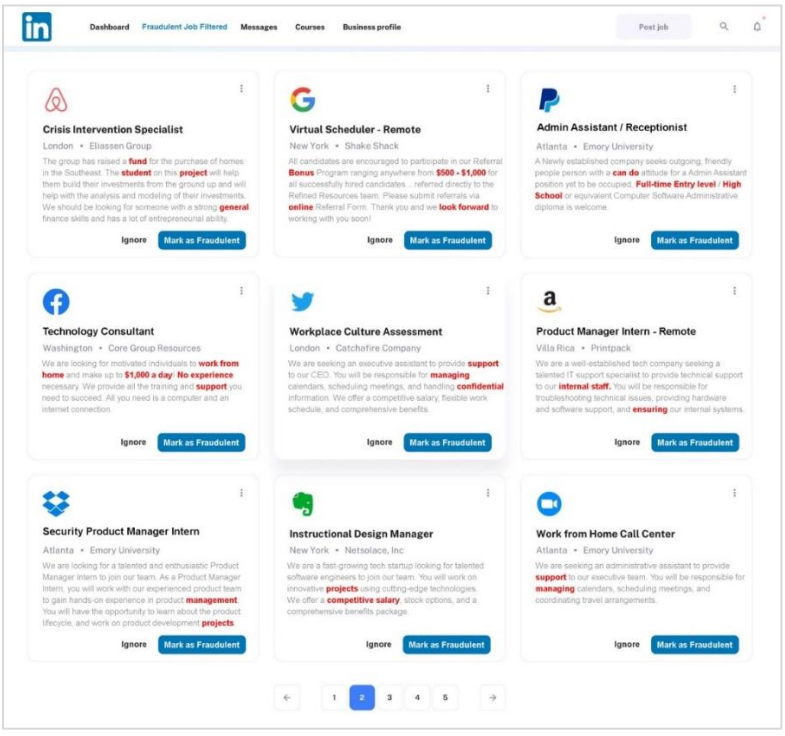


Fig. 4.2a: LinkedIn's IntelliLink Admin Dashboard showing fraudulent jobs filtered.

Example to Explain the Integration

For example, Recruiter Alex attempts to scam young and inexperienced job seekers with postings containing keywords like "project", "entry-level", and "remote". IntelliLink, integrated into LinkedIn's Talent Solutions, flags these postings and employees quickly identify them as fraudulent. Thus, Recruiter Alex's job postings were unlisted, and his account got suspended, protecting gullible job seekers from scams.

Click here to enter text.

4.2.2 Industry & Skill Demand Forecasting Integration

Our time series model for forecasting future industry and skill demand can be integrated into **LinkedIn Learning**, which provides online courses and training programs for professionals looking to acquire new skills and advance their careers *(Fig. 4.2b)*.

By forecasting the most required skills in the future, LinkedIn can help its users by recommending the most relevant courses and training programs that will benefit job seekers. LinkedIn Learning can even use this information in their process of creating new courses and training material, thus improving the relevancy of their content. LinkedIn can also consider feedback from employers to supplement the forecasting model, and even invite different companies to create and upload their own training courses.

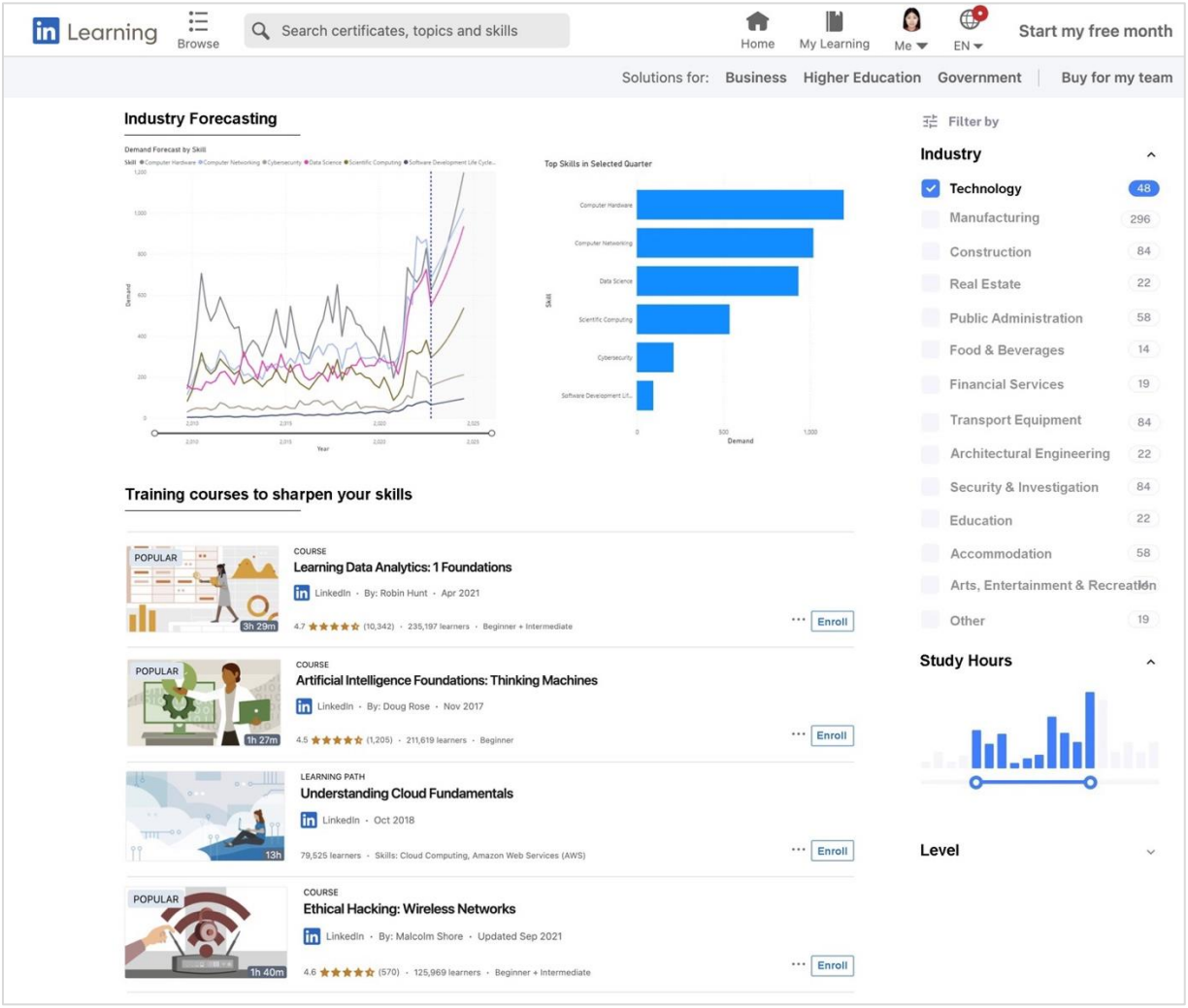


Fig. 4.2b: Platform interface showing skill demand forecast in Technology industry by year, with recommended courses.

For example, a soon to be graduating student who aspires to work in the technology industry may go on to LinkedIn Learning and filter the courses based on the industry that he or she wishes to work in. The system then filters out the most important skills in that specific industry. In our case, the forecasted top 6 most sought-after skill in the technology industry in the next 2 years are Computer Hardware, Computer Networking, Data Science, Scientific Computing, Cybersecurity and Software Development. LinkedIn Learning can thus use this information to promote courses that focus on these skills, such as “Learning Data Analytics: 1 Foundations” and “Artificial Intelligence

Click here to enter text.

Foundations: Thinking Machines". This would help the student narrow down on the courses that will be most beneficial for his or her future career.

4.2.3 Passive Job Seeker Detection Integration

The passive job seeker detection model can be integrated into LinkedIn's **Premium Subscription** segment, where recruiters can subscribe to premium features to streamline the talent search process *(Fig. 4.2c)*. With the passive jobseeker detection function, the recruiter can access to a pool of highly experienced and qualified employees who are not actively seeking employment but remain open to new opportunities. By leveraging this feature, recruiters can tap into a broader pool of talent, saving time and resources by avoiding the need to sift through irrelevant resumes.

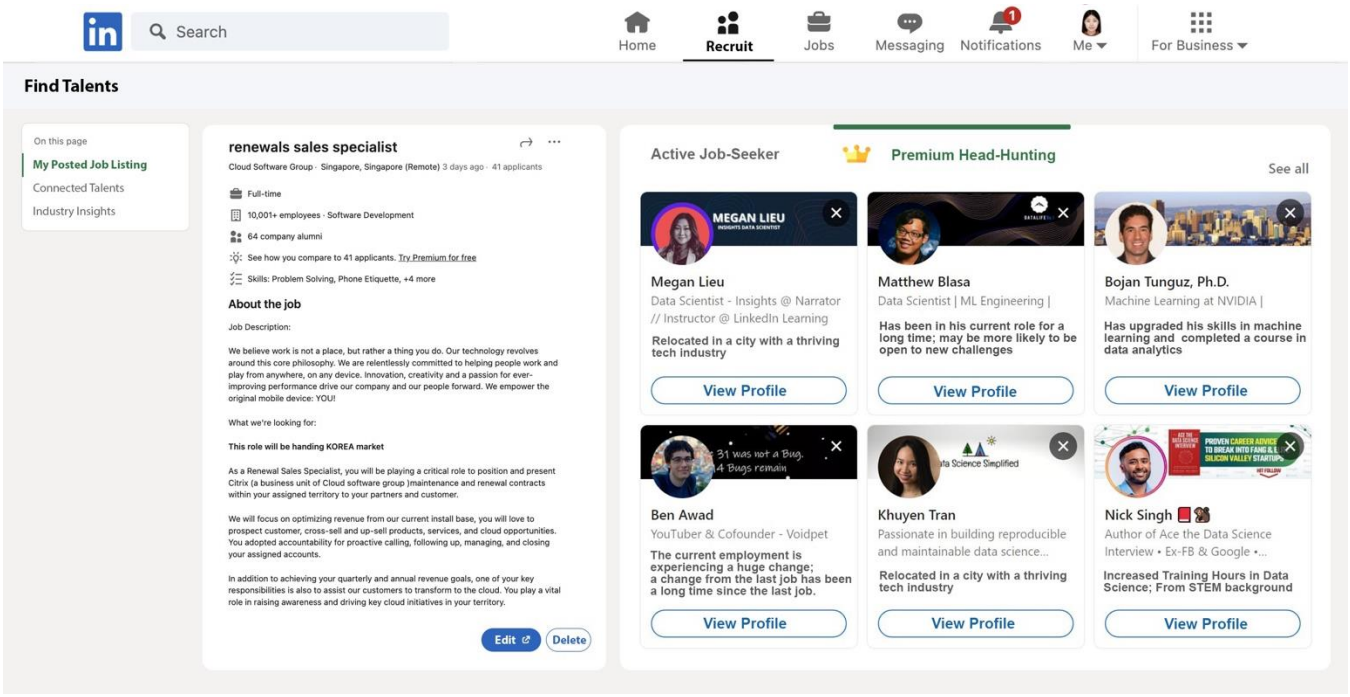


Fig. 4.2c: Passive Job Seeker Detection Integration into LinkedIn

For example, suppose a data scientist recruiter is looking for a highly qualified candidate with extensive experience for a manager position. In that case, our head-hunter feature can provide a list of passive job seekers willing to change jobs. Furthermore, our model provides an in-depth analysis report that highlights factors contributing to a candidate's potential to change jobs, such as their location, skills, experience, and more. This enables recruiters to make informed decisions and reach out to the most promising candidates, maximizing their chances of success.

For instance, our analysis may indicate that a candidate based in a less developed city with fewer job opportunities tends to switch jobs frequently to find better offers. This insight enables recruiters to understand the candidate's motivations and provide a more tailored offer, increasing the likelihood of a successful hire.

The valuable data can be collected when recruiters reach out to identified passive jobseekers, which will be fed back into the model to improve its accuracy and reliability. For instance, if a candidate responds negatively to a recruiter's message, this information can be used to update the model's understanding of what constitutes a true passive job seeker. Alternatively, if the candidate accepts the offer, this can be used as data to refine the model's predictions of what kinds of candidates are most likely to be successful hires. By continuously incorporating feedback data, the model can become more accurate and useful over time, improving the recruitment process for LinkedIn's Premium subscribers.

5. Benefits of IntelliLink

5.1 Improving the LinkedIn Recruitment Experience

By enhancing security, driving innovation, and increasing effectiveness on LinkedIn in the recruitment process, IntelliLink improves LinkedIn recruitment experience for both recruiters and job seekers *(Fig. 4.1a)*. This is achieved because IntelliLink addresses the common recruiting pain points (e.g.: scams and mismatch in skills) and streamlines and secures the recruitment process on LinkedIn. This is especially valuable since there are over 58 million companies on LinkedIn and 57% of job seekers use LinkedIn to find new job opportunities (Shepherd, 2023). Therefore, by addressing LinkedIn's job seekers and recruiters' needs, IntelliLink improves the LinkedIn recruitment experience, ensuring that LinkedIn remains a trusted and indispensable platform in the recruitment industry.

5.2 Business Profit

5.2.1 Talents Solutions

By automatically identifying and removing fraudulent job postings, LinkedIn can maintain a high level of trust and credibility among its users. This ensures that LinkedIn retain its legitimacy and brand name in the dynamic job market where scams are being more creative. Thus, this encourages more businesses and recruiters to use LinkedIn's Talent Solutions platform to find and hire talent, which in turn lead to more profits for LinkedIn. This is especially important since LinkedIn's Talents Solutions division is LinkedIn's primary source of revenue.

5.2.2 LinkedIn Learning

With IntelliLink's industry and skills demand forecasting, LinkedIn Learning can provide more targeted recommendations to users. Thus, LinkedIn can provide more value to users and increase the click through rate on the platform. Furthermore, LinkedIn Learning can enhance its competitiveness against other online learning platforms by creating new courses according to the forecasted industry needs. Since there is a huge untapped user base on LinkedIn as only 27 millions of 900 million members uses LinkedIn Learning (LinkedIn, n.d.), LinkedIn can generate more profits from their untapped user base at relatively low cost with IntelliLink and LinkedIn Learning.

The synergy between its LinkedIn Learning and its Talents Solutions segment can also be maximised. As more LinkedIn users start using LinkedIn Learning, and publishing their course certificates on their profiles, more employers will start using these certificates to filter out potential candidates. This creates a virtuous cycle where LinkedIn Learning becomes a new industry standard for legitimizing and verifying job seekers' skill sets. Therefore, providing LinkedIn with a unique edge over its competitors and increases its profits due to the increased in users.

5.2.3 Premium Subscriptions

IntelliLink's passive job seeker detection makes LinkedIn's premium subscription service more attractive for companies seeking highly skilled and qualified employees. Augmenting the daily recommendations feature in LinkedIn's Recruiter Lite subscription with IntelliLink's passive job seeker identification will increase the rates of successful hiring highly qualified job candidates. By marketing this added feature from IntelliLink, subscription rate of Recruiter Lite will increase greatly, thus increasing LinkedIn's profits.

5.3 Boost LinkedIn's Work Efficiency

IntelliLink streamlines LinkedIn's operations by automatically flagging suspicious job postings and forecast industry and skills demands. This reduces the time and effort required for manual reviews and analysis to identify fraudulent job postings and forecast industry trends frequently. Hence, resulting in a more efficient resource usage. As LinkedIn's user base grows rapidly, IntelliLink's scalability ensures that LinkedIn can remain efficient despite the rapid increase in job listings, job seekers, and recruiters.

6. Conclusion

6.1 Limitations and Concerns

6.1.1 Potential Bias

Machine learning models, like those used in IntelliLink, may inadvertently introduce biases that could affect the fairness and accuracy of their predictions. For example, people from certain backgrounds may have a higher chance of being visible to recruiters due to unintentional biases in the models. Similarly, some groups of passive job seekers might be overlooked even though they could be the perfect fit for a job. Additionally, in the case of fraudulent job listing classification, the model might only identify certain types of scams while missing other types.

To mitigate these biases, it is crucial for LinkedIn to constantly collect and update data, as well as to ensure that a diverse and representative dataset is used for model training. Regular user feedback and monitoring can help identify and address potential biases in IntelliLink too.

6.1.2 Privacy Concerns

A limitation of using analytics for identify passive job seekers is the privacy concerns it raises. LinkedIn, like many other social media platforms, collects a vast amount of personal data from its users. There is a risk that these data could be misused, leading to potential discrimination, bias, or other unethical practices. Therefore, to address these concerns, LinkedIn must only collect necessary data for legitimate prediction purposes, ensuring that there is transparency in data collection and usage, as well as compliance with data privacy regulations.

6.1.3 Inaccurate User Information

While LinkedIn possesses a wealth of data, some of it may be inaccurate due to users entering incorrect information or failing to update their profiles with their latest information regularly. This can limit the accuracy of IntelliLink's analytical models and potentially affecting the reliability of the passive job seeker recommendation system and the fraudulent job listings classification. To mitigate this problem, LinkedIn can implement measures such as providing more specific and well-phrased instructions for user input, especially for important features that may influence the model's accuracy. For example, having a guide for filling up the "Experience" and "About Me" section of the user profile. Additionally, outlier detection can be employed to filter out information that are potentially inaccurate. By employing these strategies, LinkedIn can ensure that the data collected and used is accurate, thus enhancing the performance of IntelliLink's models.

6.2 Further Considerations

6.2.1 Enhancing Model Accuracy

(a) Fraudulent Job Listing Prediction Model

For the fraudulent job listing prediction model, other predictor variables in the dataset (e.g.: telecommuting, has_company_logo, has_questions, character_count, word_count,sentence_count) beside the "processed_text" variable can be included for model training and prediction. The added variables might be important in predicting fraudulent job postings.

Additionally, hyperparameter tuning can be done to maximise the models' capabilities and achieve higher accuracy levels.

Furthermore, an alternative word embedding technique, Word2Vec, which captures semantic information can be utilized to better represent the text data in a structured format.

(b) Industry Trend Forecast Model

The industry and skill forecasting can be improved by using more data sources. Currently, the model simply uses the historical industry demand to forecast future demand. However, there are exogenous variables such as GDP,

unemployment rate, interest rate, and financing activities for various segments that are not sufficiently taken into consideration for forecasting. Therefore, more sophisticated time series models that are able to handle exogenous variables can be used. Furthermore, exponential smoothing can be used instead in derivative estimation to weigh the recent trend more heavily. Kalman filtering can also be used over a rolling average to improve estimation accuracy.

(c) Job Seeker Prediction Model

LinkedIn's extensive user database offers various features that can enhance the model's predictability. For instance, the model can use the user's active index, which identifies users who frequently engage with industry news and are more receptive to new opportunities. Another valuable feature is the completeness of user profiles, as frequent updates indicate a willingness to engage with potential employers. By incorporating more pertinent and useful data, the model is anticipated to boost its performance and provide greater transparency in its predictions.

6.2.2 Localising Data to LinkedIn

To make IntelliLink's machine learning models more tailored to LinkedIn, they can be re-trained using LinkedIn-specific data. This approach enhances the models' prediction accuracy and ensures that the results are more reliable and applicable to LinkedIn's unique context. The process of re-training the models can be done by following the machine learning pipeline established in the source code, but with input data from LinkedIn.

6.2.3 Improving Users' Trust

The machine learning models used in IntelliLink are "black boxes" that offer limited insights into the reasoning behind individual predictions. This makes it difficult for the users to trust the predictions, potentially resulting in them making decisions that contradict the model's recommendations. Thus, to improve users' trust, a model interpreter, such as SHAP (SHapley Additive exPlanations), can be used to shed light on what contributes to the final model prediction for each specific prediction. By revealing the inner workings of the model predictions and offering a clearer understanding of the prediction process, users will trust and use IntelliLink more.

6.2.4 Detecting and Removing Fake Profiles and Information

To further increase IntelliLink's accuracy and reliability, the problem of fake profiles and inaccurate information on LinkedIn can be tackled. Implementing a fake profile/information detection system can help mitigate the impact of false or misleading data on LinkedIn, as well as ensure that IntelliLink's models perform accurately and reliably. Some features that can be used to train the model include account activity, the authenticity of the profile picture, connections, and whether the profile information aligns with industry norms.

By proactively identifying and flagging suspicious profiles, LinkedIn employees can verify or remove them. Thus, reducing the negative impact of fake profiles and information on intelliLink's performance, as well as maintaining the LinkedIn's credibility.

6.3 Ending Remarks

In conclusion, IntelliLink is a ground-breaking solution that offers numerous advantages and enhancements for LinkedIn's platform. By integrating IntelliLink's suite of machine learning models, LinkedIn can solidify its position as the go-to platform for both job seekers and recruiters. The benefits of enhanced recruiting experience for jobseekers and recruiters, the increased business profits, the improved efficiency, and the strengthened recruitment security create a compelling case for adopting IntelliLink.

Imagine a future where job seekers can confidently pursue opportunities without fear of fraudulent LinkedIn job listings and are empowered with the knowledge of trending skills that ensures the success of their career. At the same time, recruiters can tap into a broader talent pool comprising of passive job seekers and have higher quality hires and higher successful placement rates. IntelliLink makes this future a reality.

7. References

About linkedin. About LinkedIn. (n.d.). Retrieved February 26, 2023, from https://about.linkedin.com/

Boksic, B. (2022, July 6). *What is a skills mismatch and how do you solve it?* Vervoe. Retrieved March 30, 2023, from https://vervoe.com/skills-mismatch/

Brownlee, J. (2020, January 14). *Why is imbalanced classification difficult?* Machine Learning Mastery. Retrieved October 28, 2022, from https://machinelearningmastery.com/imbalanced-classification-is-hard/

Cedefop. (2012). Skill mismatch: The role of the enterprise. European Centre for the Development of Vocational Training Research Paper, 21.

Chua, N. (2022). $227.8m lost to top 10 scams in first half of 2022, as overall crime rises by 36%. The Straits Times. Retrieved February 26, 2023, from https://www.straitstimes.com/singapore/courts-crime/2278m-lost-to-top-10-scams-in-first-half-of-2022-as-overall-crime-rises-by-36

Cue. (2023, March 17). *Retrenchments in s'pore doubled in Q4 of 2022, but fewer jobs cut in tech than expected*. The Straits Times. Retrieved March 30, 2023, from https://www.straitstimes.com/singapore/jobs/retrenchments-in-q4-2022-double-of-q3-mom-report#:~:text=The%20corresponding%20figure%20for%20Singaporeans,cent%20to%203%20per%20cent.&text=As%20a%20whole%2C%20unemployment%20fell,2.7%20per%20cent%20in%202021.

Graham, D. (2020) Employment-related scams are on the rise: Learn how to protect yourself, Forbes. Forbes Magazine. Available at: https://www.forbes.com/sites/dawngraham/2020/08/04/employment-related-scams-are-on-the-rise-learn-how-to-protect-yourself/?sh=2710ea7877e5 (Accessed: February 14, 2023).

HRKatha, K.K.| H.R.K. et al. (2022) Why is the job market becoming so volatile?, HR Katha. Available at: https://www.hrkatha.com/features/why-is-the-job-market-becoming-so-volatile/ (Accessed: February 14, 2023).

Jones, S. (2021). Drawbacks of traditional hiring methods and why they no longer work. Uplers. Retrieved February 16, 2023, from https://www.uplers.com/blog/drawbacks-of-traditional-hiring-methods/

Khorram, Y., & Zamost, S. (2022, June 17). *FBI says fraud on linkedin a 'significant threat' to platform and consumers*. CNBC. Retrieved April 1, 2023, from https://www.cnbc.com/2022/06/17/fbi-says-fraud-on-linkedin-a-significant-threat-to-platform-and-consumers.html

Pahwa, A. (2023, February 15). How LinkedIn Makes Money? LinkedIn Business Model. Feedough. https://www.feedough.com/how-linkedin-makes-money/

Petrone, P. (2020). 7 of the biggest problems recruiters face (and how to overcome them). LinkedIn. Retrieved February 16, 2023, from https://www.linkedin.com/business/talent/blog/talent-acquisition/biggest-problems-recruiters-face-and-how-to-overcome-them

Recruiting active vs. passive candidates. LinkedIn. (2013). Retrieved February 16, 2023, from https://www.linkedin.com/business/talent/blog/talent-acquisition/recruiting-active-vs-passive-candidates

Rella, E. (2022). Woman Gets Scammed with Fake LinkedIn Job Posting -- Here Are the Red Flags. Entrepreneur. https://www.entrepreneur.com/business-news/linkedin-job-scams-are-rampant-here-are-the-signs/434799

Ruby, D., & About The Author Daniel Ruby Content writer with 10+ years of experience. I write across a range of subjects. (2023, February 28). *69+ layoff statistics for 2023 (Latest Data & Future Trends)*. Demand Sage. Retrieved March 30, 2023, from https://www.demandsage.com/layoff-statistics/

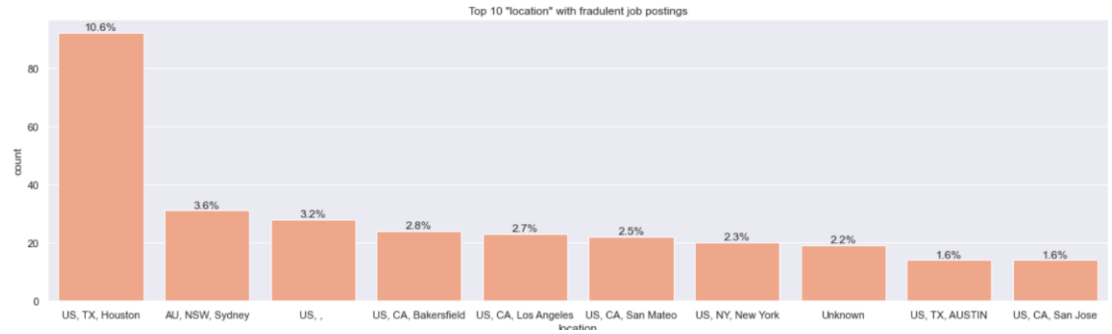
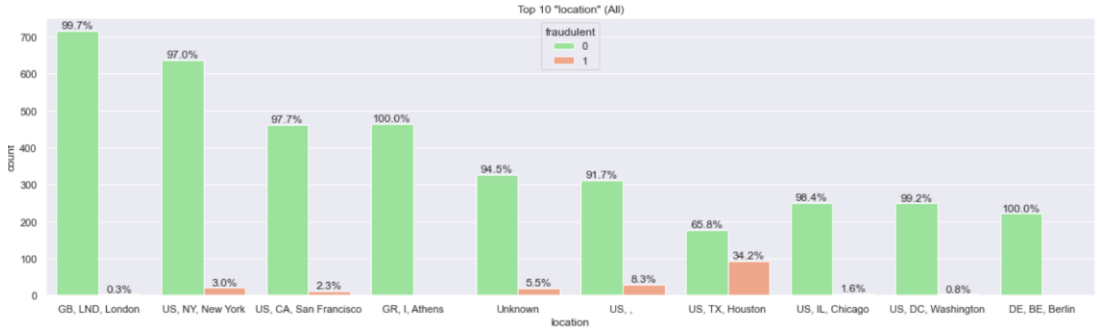
Shepherd, J. (2023, February 23). 40 essential LinkedIn statistics you need to know in 2023. The Social Shepherd. https://thesocialshepherd.com/blog/linkedin-statistics

Click here to enter text.

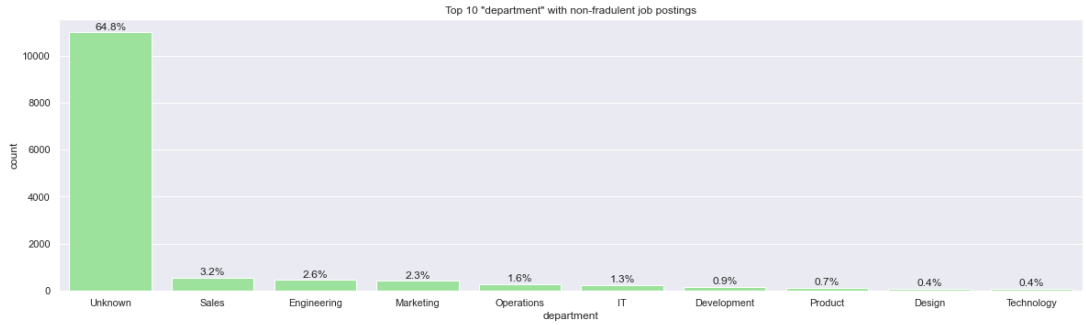
Appendix A – Fraudulent Job Listings Prediction

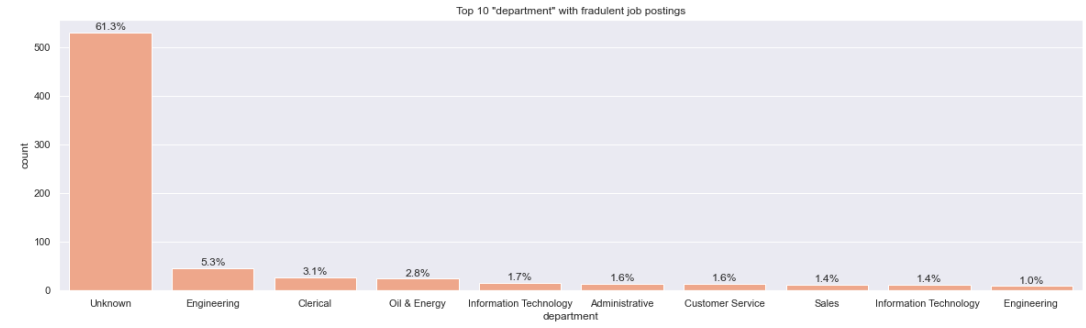
A1. More Exploratory Data Analysis

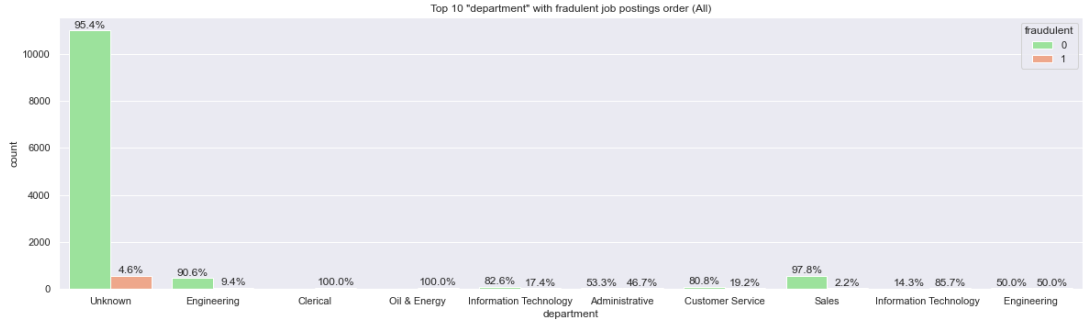
This section contains more insights from the exploratory data analysis for the fraudulent job listings prediction. If you are interested to see the whole EDA process and all the insights obtained, please visit: https://github.com/leileijng/bc2407-linkedin/blob/main/Job%20Scam%20Detection/exploratory-data-analysis.ipynb

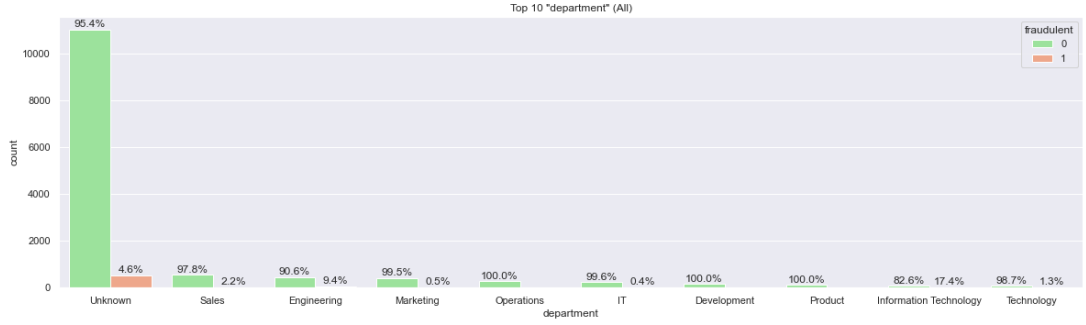
Diagram	Insights
	**Location** Most of the job listings are from US, TX, Houston (10.6%)
	**Location** Job listings from US, CA, Bakersfield and US, TX, AUSTIN have a very high chance of being a scam (100% and 92.3%)
	**Location** Job listings from GB, LND, London has an extremely low chance of being a scam (99.7%). Job listings from US, NY, New York has an extremely low chance of being a scam too (97.0%)

Click here to enter text.









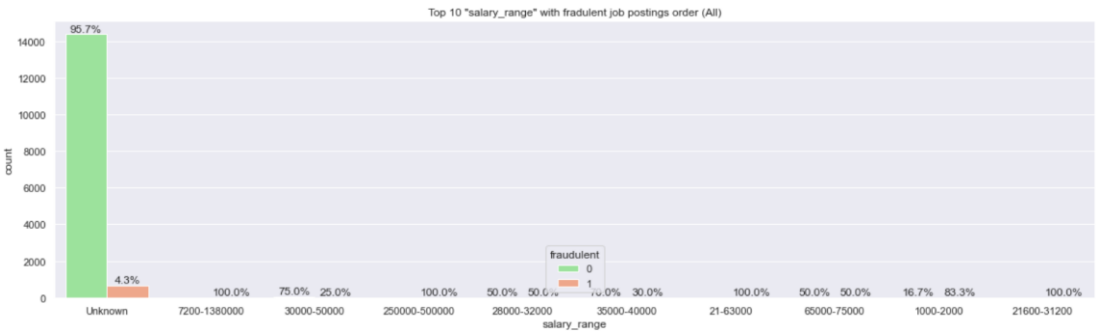
Department

Most of the job listings have Unknown departments regardless of whether it is fraudulent or not (~60%)

A sizable number of job listings have Engineering department (5.3%) (excluding Unknown department)

Job listings with Clerical and Oil & Energy departments have a very high chance of being a scam (both 100%)

Job listings with Unkwown departments suprisingly has a low chance of being a scam (95.4%)

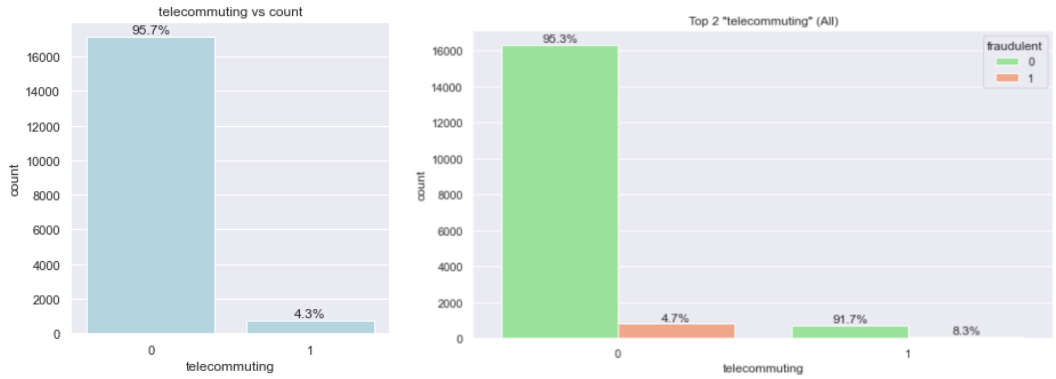
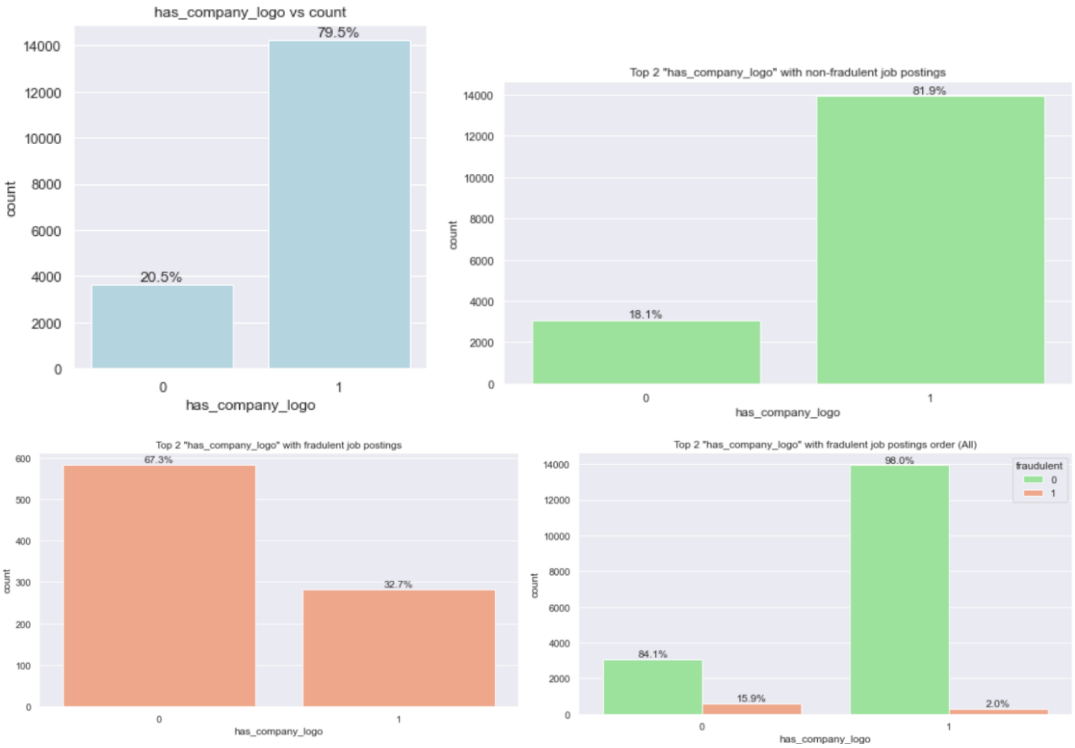


Salary Range

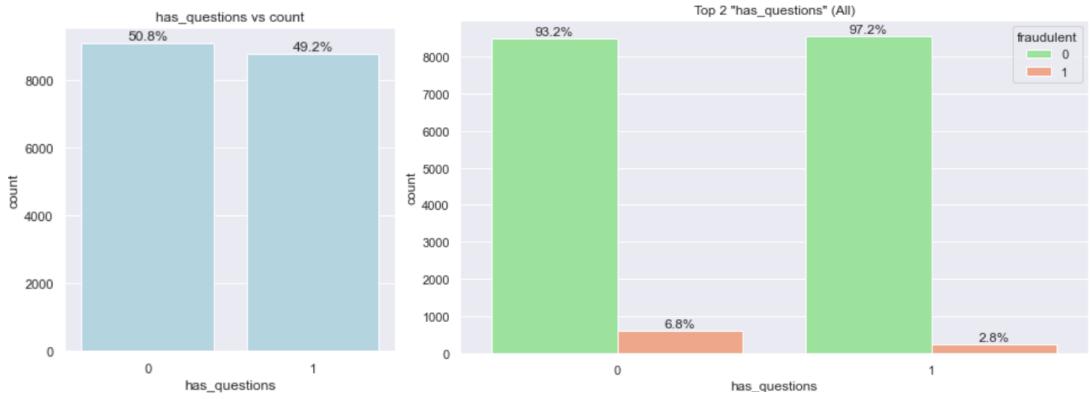
Job listings with large salary range of 7200-1380000 and 250000-50000 salary range have a very high chance of being a scam (both 100%)

Job listings with Unknown and 0-0

Click here to enter text.

	salary range surprisingly has a low chance of being a scam (95.7% and 97.2%)
	**<u>Telecommuting</u>** Most of the job listings does not have telecommuting (95.7%) A job listing that has telecommuting is about 2 times more likely to be a scam compared to a job listing without telecommuting (8.3% vs 4.7%)
	**<u>Company Logo</u>** Most of the job listings has a company logo (79.5%) Most of the job listings that are not fraudulent has a company logo (67.3%) Most of the job listings that are fraudulent does not have a company logo (81.9%) A job listing that does not have a company logo has a significantly higher chance of being a scam compared to that that has a company logo (15.9% vs 2.0%)

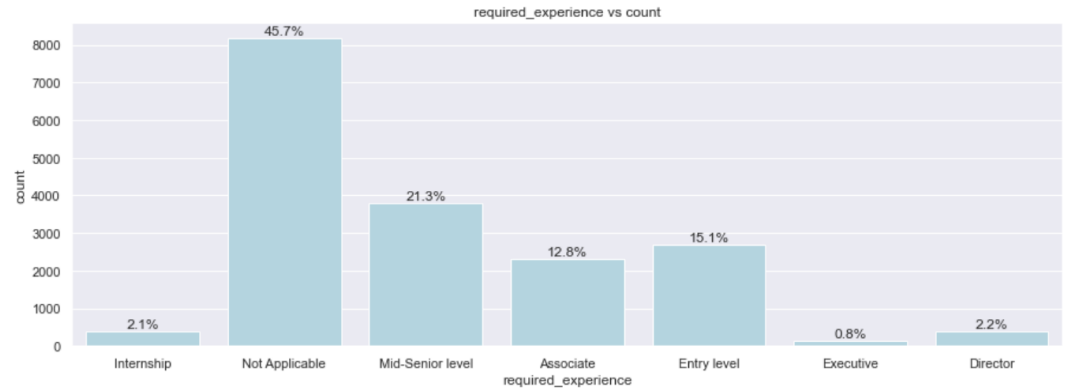
Click here to enter text.

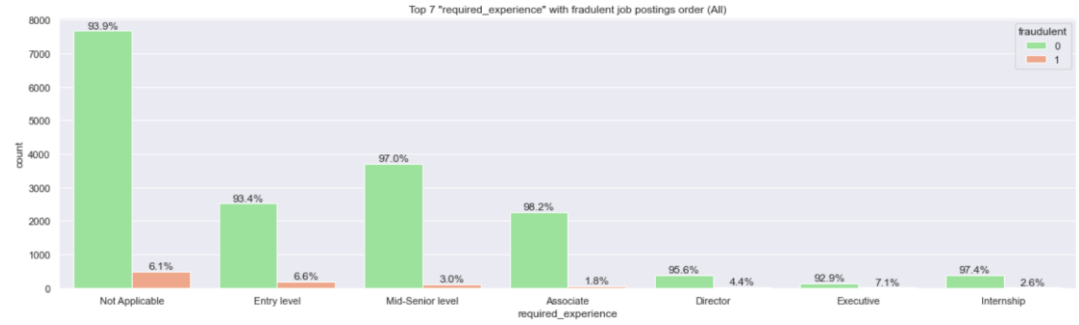


Has Questions

There is an even distribution of has_question for fraudulent and non-fraudulent job listings.

A job listing that does not have question is about 2.5 times more likely to be a scam compared to that that has_question (6.8% vs 2.8%)





Required Experience

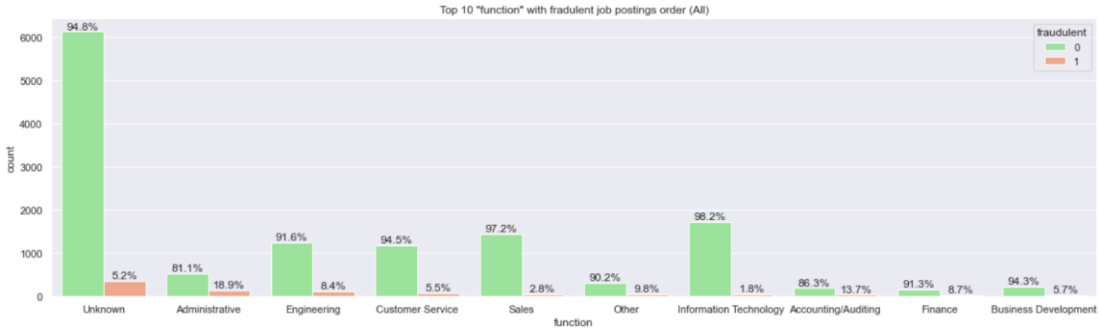
Most of the job listings do not require experience (i.e. Not Applicable) (45.7%)

Job listings with with Executive, Entry level, and Not Applicable have a higher chance of being a scam compared to other levels of required_experience

Industry

Job listings from Oil & Energy or Accounting industries have a very high chance of being a scam (28.0% and 25.8%)

Click here to enter text.

	**Function** Job listings with Administrative or Accounting/Auditing job functions have a very high chance of being a scam (18.9% and 13.7%)

Click here to enter text.

A2. Detailed Model Test Performance Results

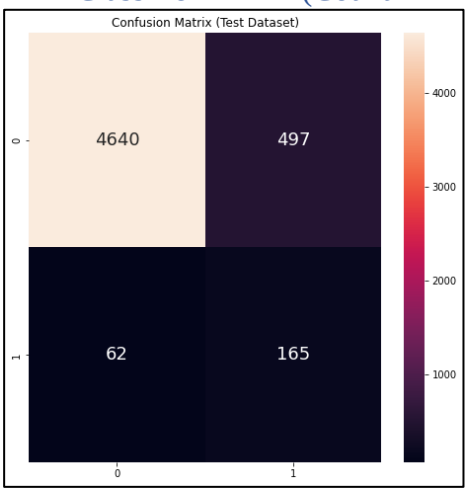
A2.1 Multinomial Naive Bayes Classifier (Count Vectorized)

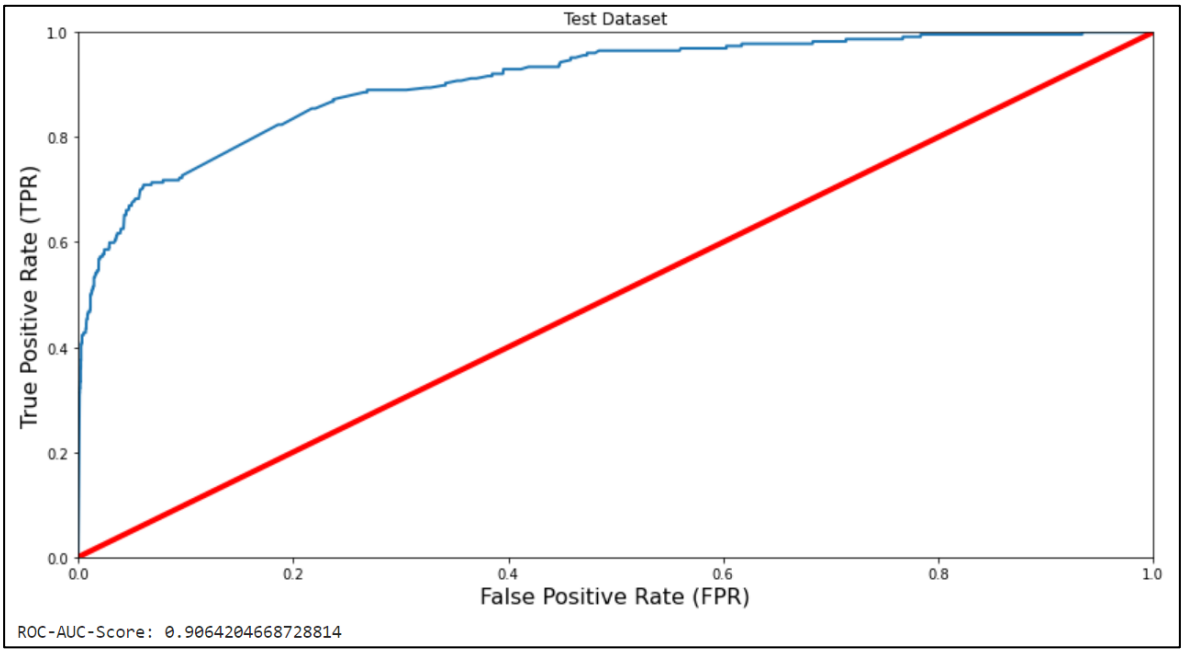
```
Accuracy score: 0.8957867263236391
              precision    recall  f1-score   support

           0       0.99      0.90      0.94      5137
           1       0.25      0.73      0.37       227

    accuracy                           0.90      5364
   macro avg       0.62      0.82      0.66      5364
weighted avg       0.96      0.90      0.92      5364

Goodness of Fit of Model (Test Dataset)
Classification Accuracy: 0.8957867263236391
True Positive Rate: 0.7268722466960352
True Negative Rate: 0.9032509246642009
False Postive Rate: 0.0967490753357991
False Negative Rate: 0.27312775330396477
```





ROC-AUC-Score: 0.9064204668728814

Click here to enter text.

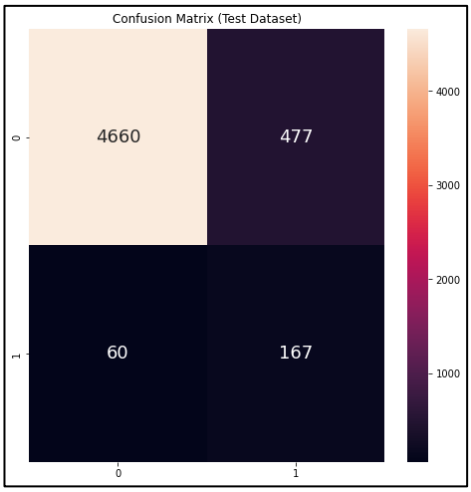
A2.2 Multinomial Naive Bayes Classifier (tf-idf Vectorized)

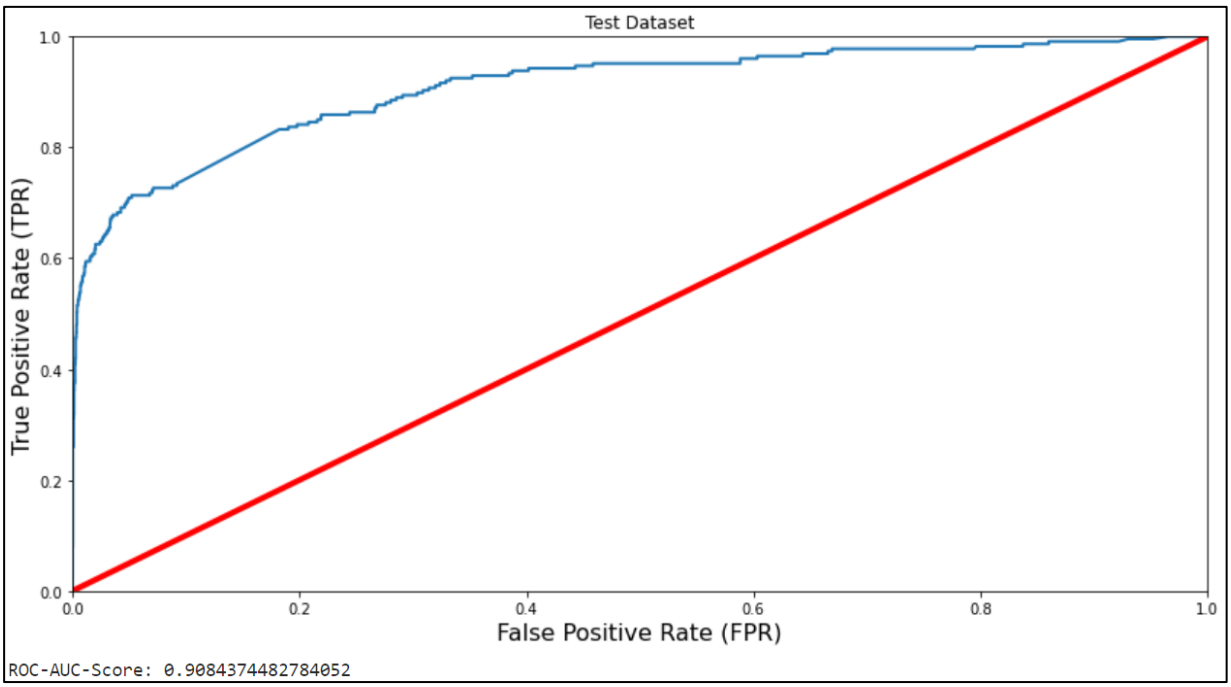
```
Accuracy score: 0.8998881431767338
              precision    recall  f1-score   support

           0       0.99      0.91      0.95      5137
           1       0.26      0.74      0.38       227

    accuracy                           0.90      5364
   macro avg       0.62      0.82      0.66      5364
weighted avg       0.96      0.90      0.92      5364

Goodness of Fit of Model (Test Dataset)
Classification Accuracy: 0.8998881431767338
True Positive Rate: 0.73568281938326
True Negative Rate: 0.9071442476153397
False Postive Rate: 0.0928557523846603
False Negative Rate: 0.2643171806167401
```





ROC-AUC-Score: 0.9084374482784052

Click here to enter text.

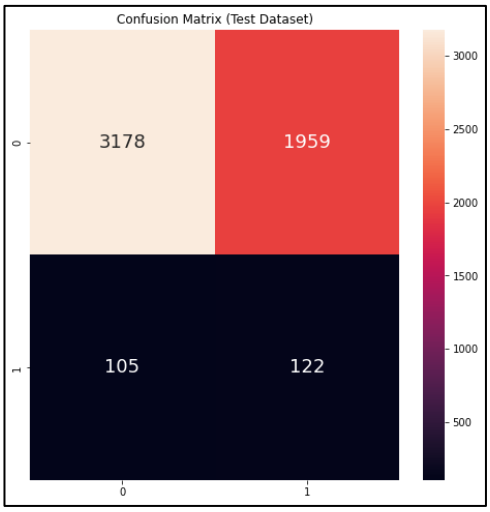
A2.3 Support Vector Classifier (Count Vectorized)

```
Accuracy score: 0.6152125279642058
              precision    recall  f1-score   support

           0       0.97      0.62      0.75      5137
           1       0.06      0.54      0.11       227

    accuracy                           0.62      5364
   macro avg       0.51      0.58      0.43      5364
weighted avg       0.93      0.62      0.73      5364

Goodness of Fit of Model (Test Dataset)
Classification Accuracy: 0.6152125279642058
True Positive Rate: 0.5374449339207048
True Negative Rate: 0.6186490169359549
False Postive Rate: 0.38135098306404513
False Negative Rate: 0.46255506607929514
```



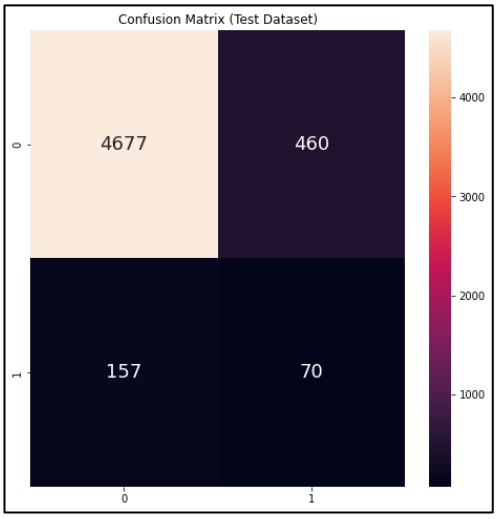
A2.4 Support Vector Classifier (tf-idf Vectorized)

```
Accuracy score: 0.8849739000745712
              precision    recall  f1-score   support

           0       0.97      0.91      0.94      5137
           1       0.13      0.31      0.18       227

    accuracy                           0.88      5364
   macro avg       0.55      0.61      0.56      5364
weighted avg       0.93      0.88      0.91      5364

Goodness of Fit of Model (Test Dataset)
Classification Accuracy: 0.8849739000745712
True Positive Rate: 0.30837004405286345
True Negative Rate: 0.9104535721238076
False Postive Rate: 0.08954642787619232
False Negative Rate: 0.6916299559471366
```



Click here to enter text.

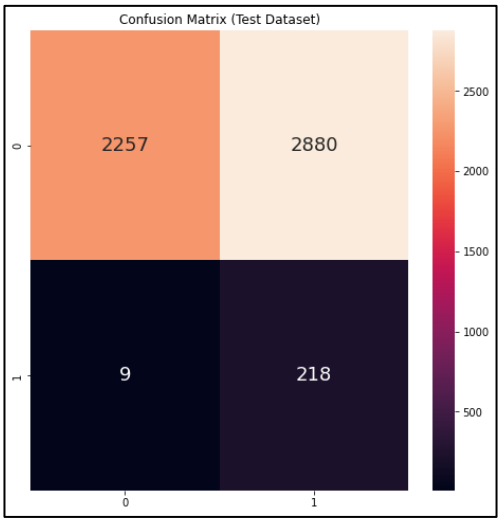
A2.5 Logistic Regression (Count Vectorized)

```
Accuracy score: 0.46140939597315433
              precision    recall  f1-score   support

           0       1.00      0.44      0.61      5137
           1       0.07      0.96      0.13       227

    accuracy                           0.46      5364
   macro avg       0.53      0.70      0.37      5364
weighted avg       0.96      0.46      0.59      5364

Goodness of Fit of Model (Test Dataset)
Classification Accuracy: 0.46140939597315433
True Positive Rate: 0.960352422907489
True Negative Rate: 0.43936149503601324
False Postive Rate: 0.5606385049639868
False Negative Rate: 0.039647577092511016
```



ROC-AUC-Score: 0.873551044979886

Click here to enter text.

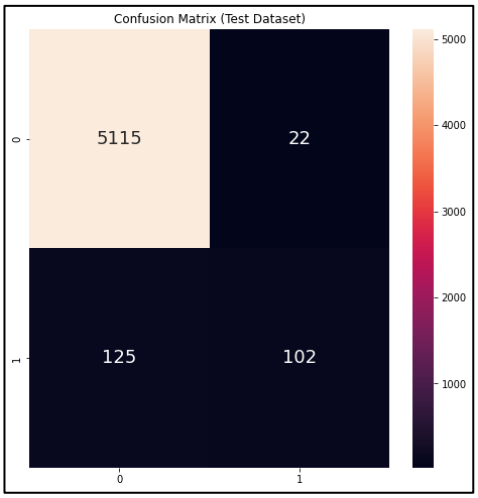
A2.6 Logistic Regression (tf-idf Vectorized)

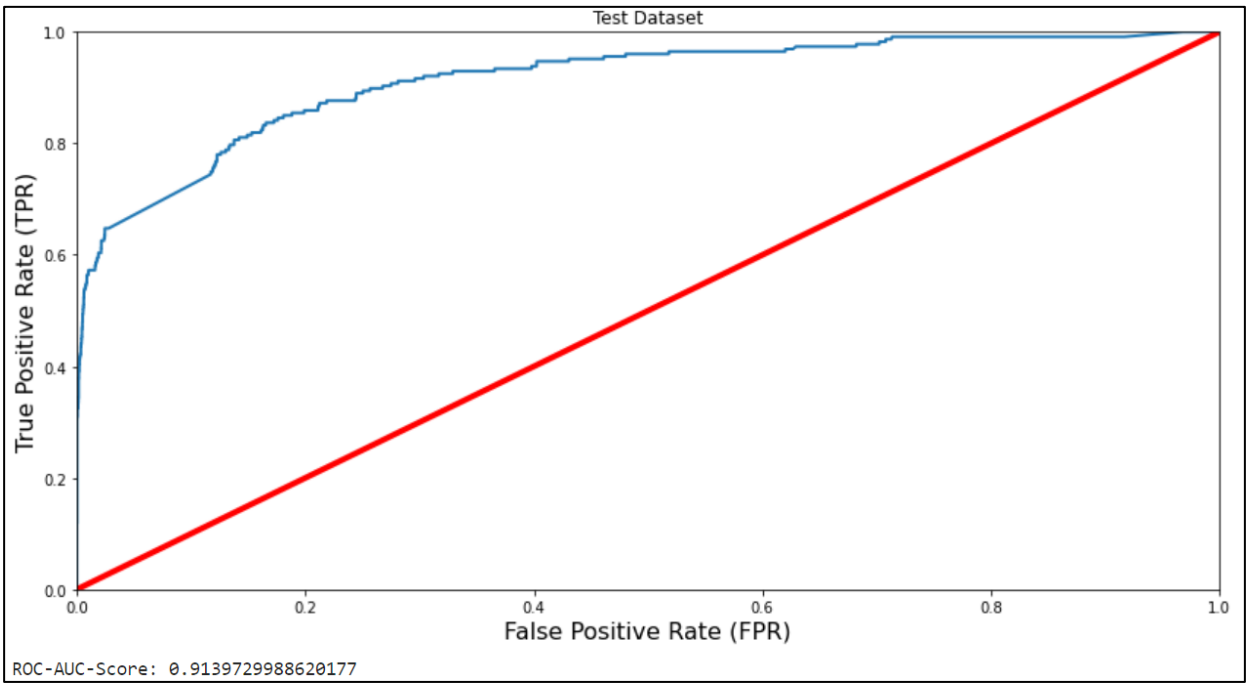
```
Accuracy score: 0.9725950782997763
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      5137
           1       0.82      0.45      0.58       227

    accuracy                           0.97      5364
   macro avg       0.90      0.72      0.78      5364
weighted avg       0.97      0.97      0.97      5364

Goodness of Fit of Model (Test Dataset)
Classification Accuracy: 0.9725950782997763
True Positive Rate: 0.44933920704845814
True Negative Rate: 0.9957173447537473
False Postive Rate: 0.004282655246252677
False Negative Rate: 0.5506607929515418
```





ROC-AUC-Score: 0.9139729988620177

Click here to enter text.

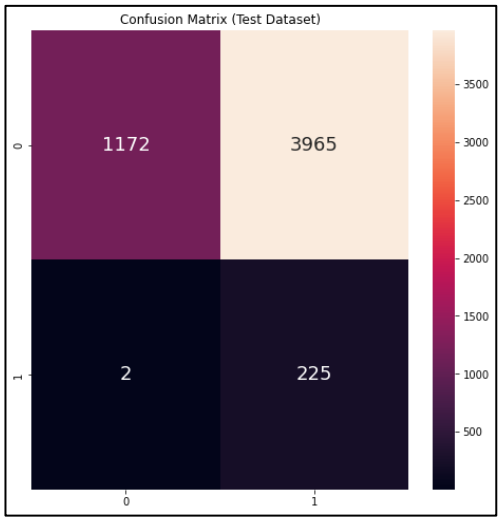
A2.7 Random Forest (Count Vectorized)

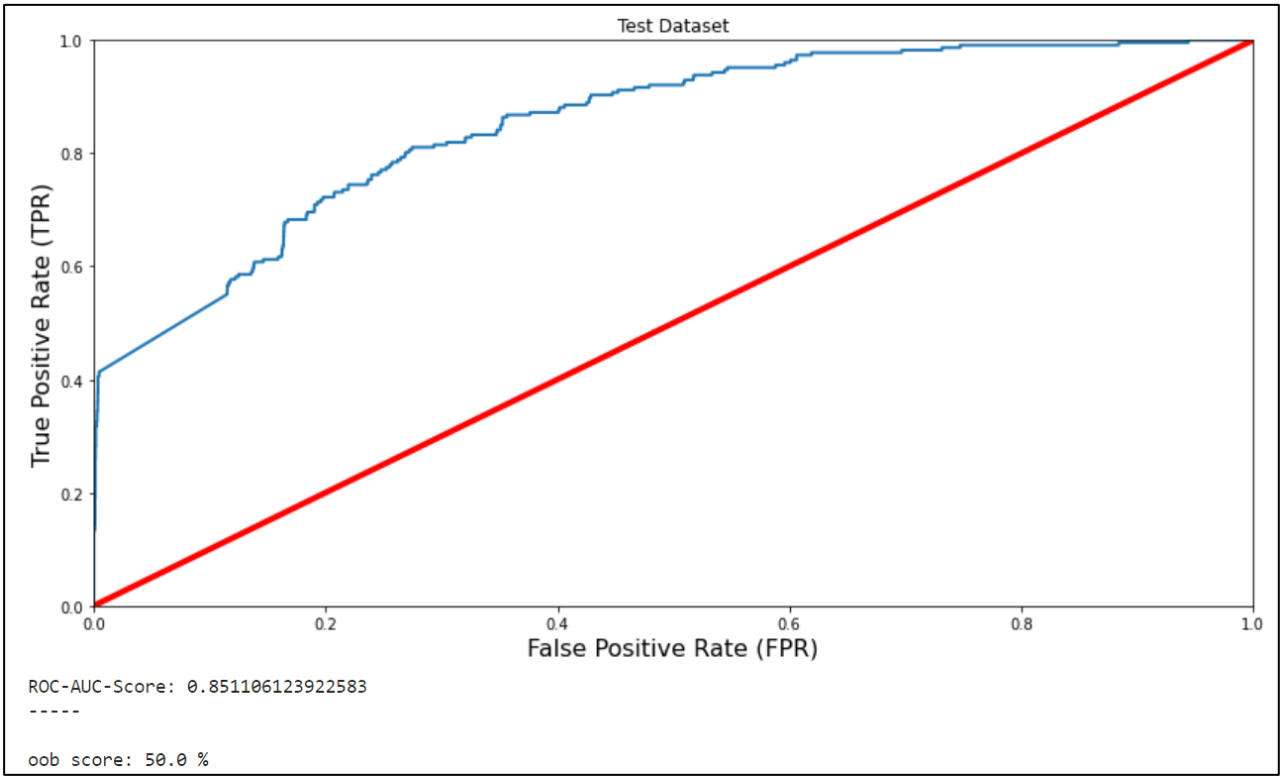
```
Accuracy score: 0.26043997017151377
              precision    recall  f1-score   support

           0       1.00      0.23      0.37      5137
           1       0.05      0.99      0.10       227

    accuracy                           0.26      5364
   macro avg       0.53      0.61      0.24      5364
weighted avg       0.96      0.26      0.36      5364

Goodness of Fit of Model (Test Dataset)
Classification Accuracy: 0.26043997017151377
True Positive Rate: 0.9911894273127754
True Negative Rate: 0.2281487249367335
False Postive Rate: 0.7718512750632665
False Negative Rate: 0.00881057268722467
```






```
ROC-AUC-Score: 0.851106123922583
-----

oob score: 50.0 %
```


Click here to enter text.

Appendix B – Industry Demand Forecasting

B1. Dataset Links

- Job Vacancy of Singapore: https://www.kaggle.com/datasets/subhamjain/job-vacancy-of-singapore-annual?select=metadata-job-vacancy-by-industry-and-occupational-group-annual.txt
- Job Vacancy of Singapore (Quarterly): https://tablebuilder.singstat.gov.sg/table/TS/M184071
- Industry Skills Needs: https://datacatalog.worldbank.org/search/dataset/0038027

B2. Target Variables

Target variables include the quarterly vacancy of the different industries in Singapore from 2006 Q2 to 2022 Q3. In total, there are 43 different columns representing different industries. Industries are also ranked in a hierarchy, for example, Goods Producing Industries is the sum of the Manufacturing and Construction industries, and Total is the sum of Goods Producing Industries and Services.

Total

Goods Producing Industries

Manufacturing

Food, Beverages & Tobacco

Paper/Rubber/Plastic Products & Printing

Petroleum, Chemical & Pharmaceutical Products

Fabricated Metal Products, Machinery & Equipment

Electronic, Computer & Optical Products

Transport Equipment

Other Manufacturing Industries

Construction

Services

Wholesale And Retail Trade

Wholesale Trade

Retail Trade

Transportation And Storage

Land Transport & Supporting Services

Water Transport & Supporting Services

Air Transport & Supporting Services

Other Transportation & Storage Services

Accommodation And Food Services

Accommodation

Food & Beverage Services

Information And Communications

Telecommunications, Broadcasting & Publishing

Click here to enter text.

- IT & Other Information Services
- Financial And Insurance Services
  - Financial Services
  - Insurance Services
- Real Estate Services
- Professional Services
  - Legal, Accounting & Management Services
  - Architectural & Engineering Services
  - Other Professional Services
- Administrative And Support Services
  - Security & Investigation
  - Cleaning & Landscaping
  - Other Administrative & Support Services
- Community, Social And Personal Services
  - Public Administration & Education
  - Health & Social Services
  - Arts, Entertainment & Recreation
  - Other Community, Social & Personal Services
- Others

Since the industries higher up the hierarchy is just the sum of its sub industries, we only need to forecast the industries at the bottom. By removing these parent industries, we are left with just 32 industries to forecast.

Food, Beverages & Tobacco

Paper/Rubber/Plastic Products & Printing

Petroleum, Chemical & Pharmaceutical Products

Fabricated Metal Products, Machinery & Equipment

Electronic, Computer & Optical Products

Transport Equipment

Other Manufacturing Industries

Construction

Wholesale Trade

Retail Trade

Land Transport & Supporting Services

Water Transport & Supporting Services

Air Transport & Supporting Services

Other Transportation & Storage Services

Accommodation

Food & Beverage Services

Telecommunications, Broadcasting & Publishing

IT & Other Information Services

Financial Services

Insurance Services

Real Estate Services

Legal, Accounting & Management Services

Architectural & Engineering Services

Other Professional Services

Security & Investigation

Cleaning & Landscaping

Other Administrative & Support Services

Public Administration & Education

Health & Social Services

Arts, Entertainment & Recreation

Other Community, Social & Personal Services

Others

Click here to enter text.

B2.1 Raw Data

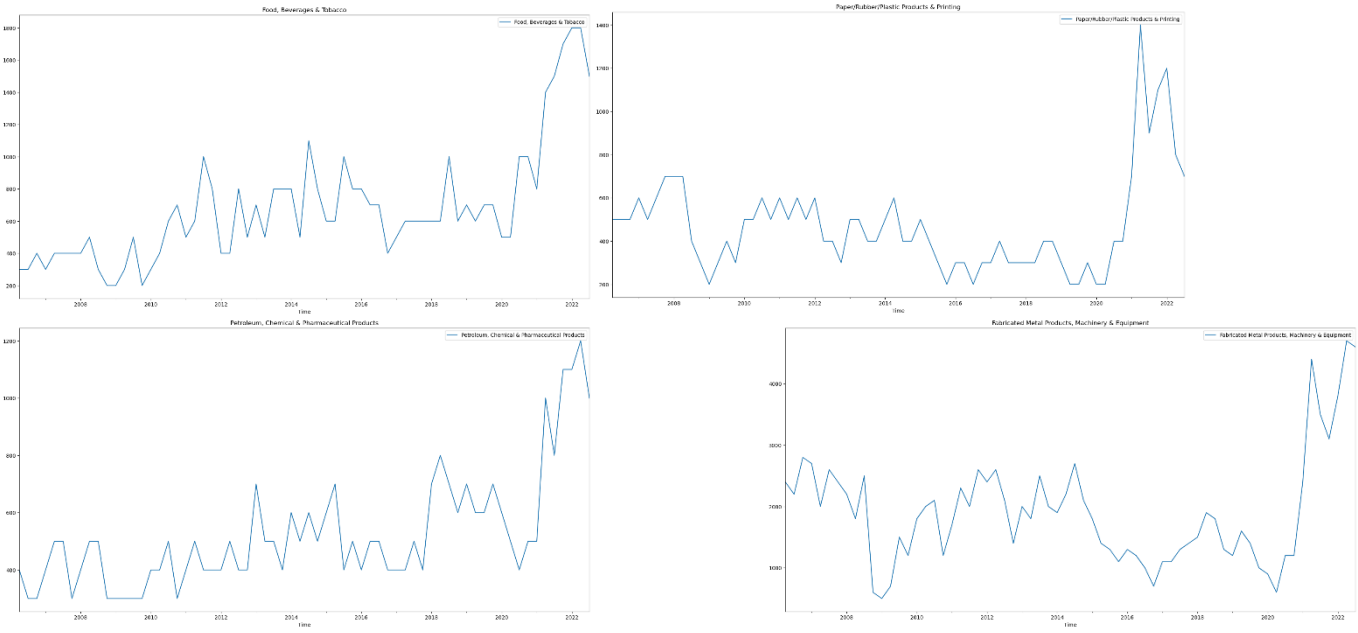
Sample of the csv file of the child industries:

	Time	Food, Beverages & Tobacco	Paper/Rubber/Plastic Products & Printing	Petroleum, Chemical & Pharmaceutical Products	Fabricated Metal Products, Machinery & Equipment	Electronic, Computer & Optical Products	Transport Equipment	Other Manufacturing Industries	Construction	Wholesale Trade	...	Architectural & Engineering Services
0	2006-04-01	300	500	400	2400	3000	2000	700	2200	1600	...	600
1	2006-07-01	300	500	300	2200	2800	1800	600	1400	1900	...	500
2	2006-10-01	400	500	300	2800	2100	1500	400	1900	2300	...	700
3	2007-01-01	300	600	400	2700	2400	1900	700	1800	2400	...	700
4	2007-04-01	400	500	500	2000	2000	2200	800	2400	2800	...	1000
...	...	...	...	...	...	...	...	...	...	...	...	...
61	2021-07-01	1500	900	800	3500	3100	1800	1800	11900	6000	...	2100
62	2021-10-01	1700	1100	1100	3100	3000	2000	1800	12400	5700	...	2400
63	2022-01-01	1800	1200	1100	3800	3100	2000	1500	12300	6600	...	2600
64	2022-04-01	1800	800	1200	4700	3600	2400	1400	11700	6800	...	3200
65	2022-07-01	1500	700	1000	4600	2600	1800	1500	9600	5900	...	2400

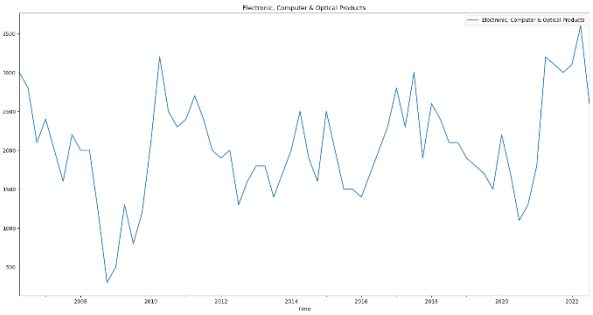
66 rows × 33 columns

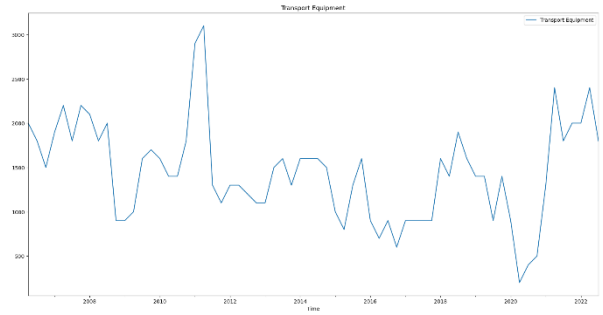
B2.2 Time Series Plots

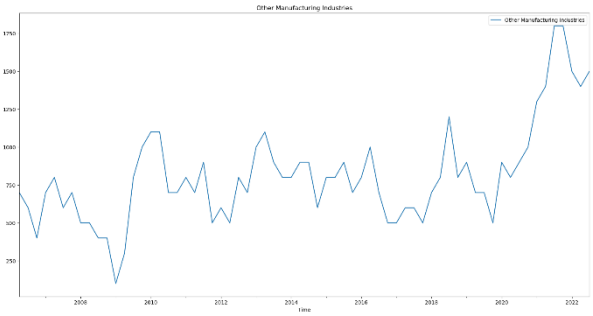
Below are the plots for each industry.

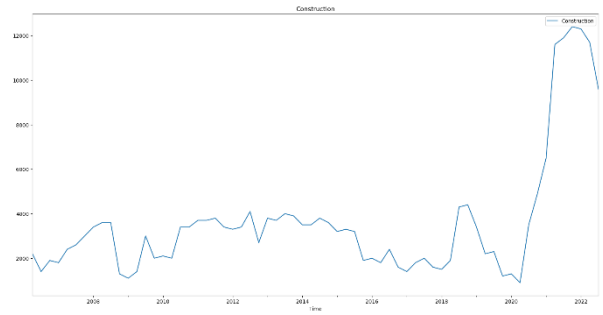


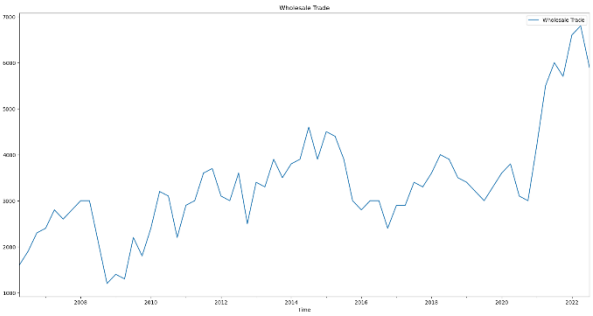
Click here to enter text.

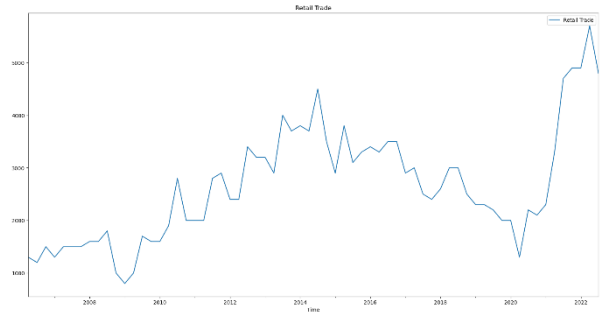


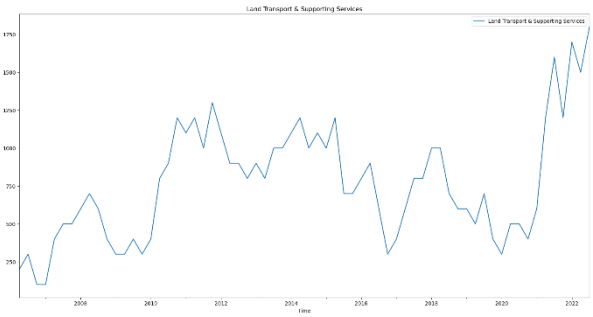


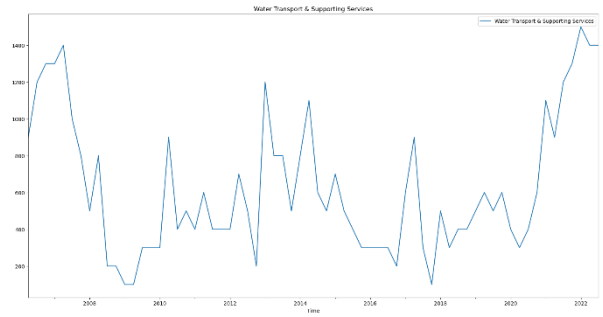


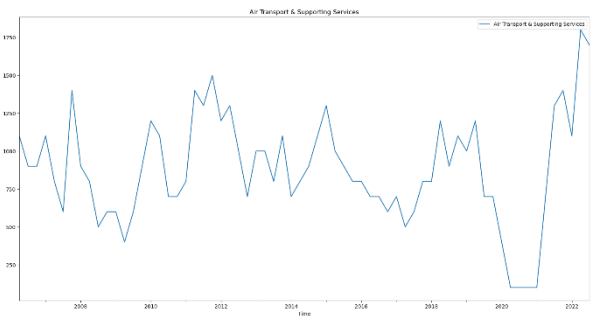


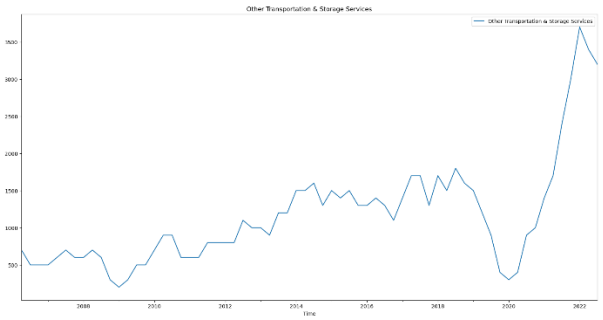




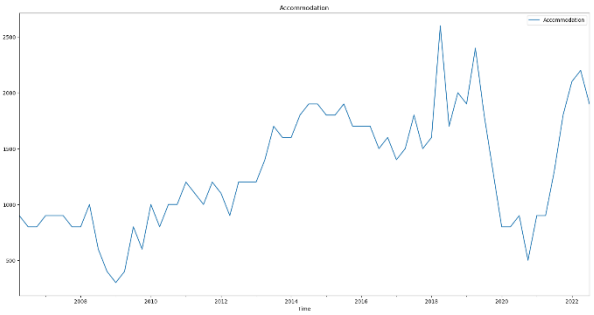


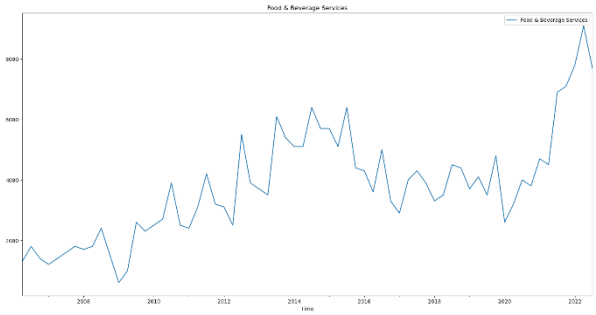


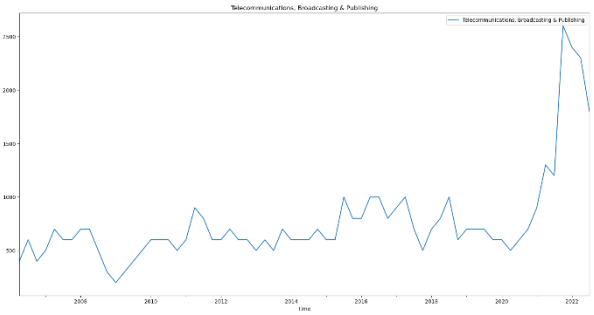


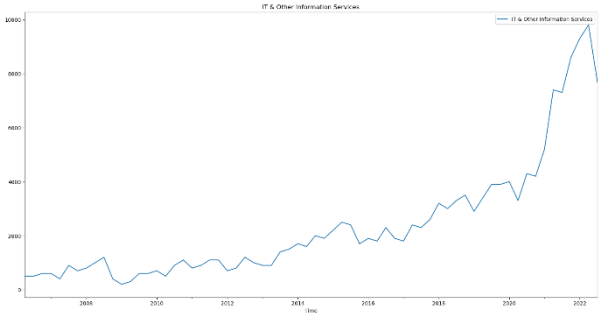


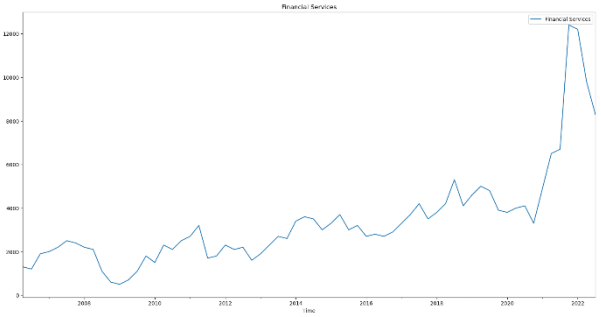
Click here to enter text.

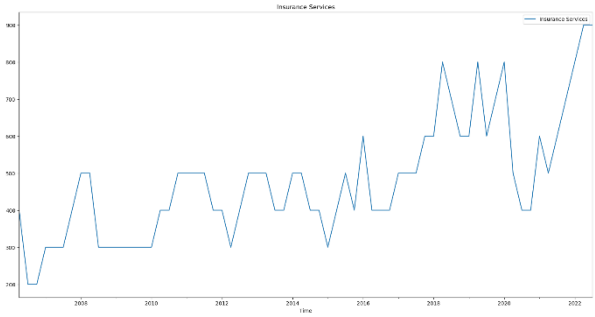


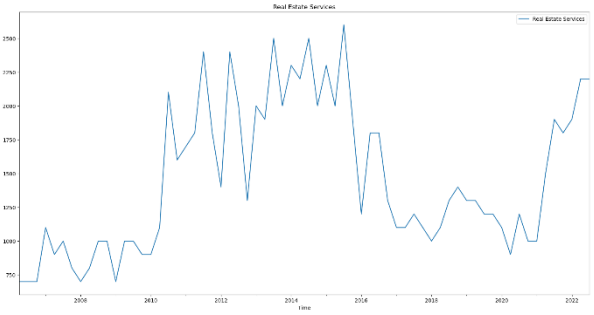


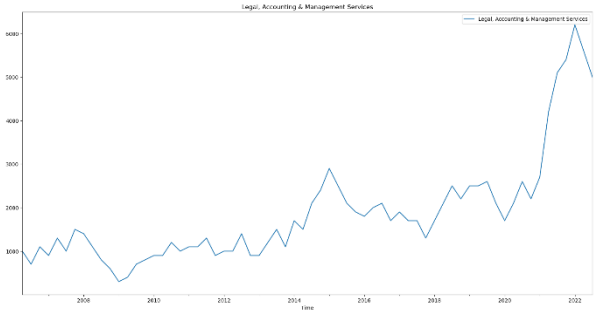


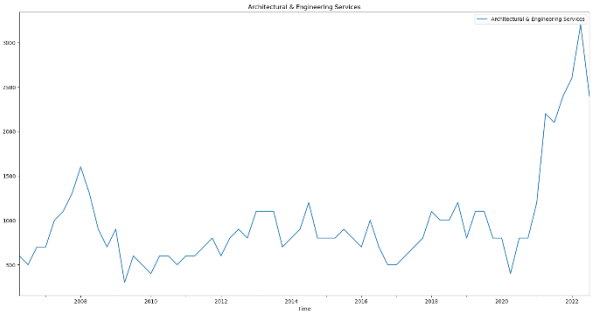


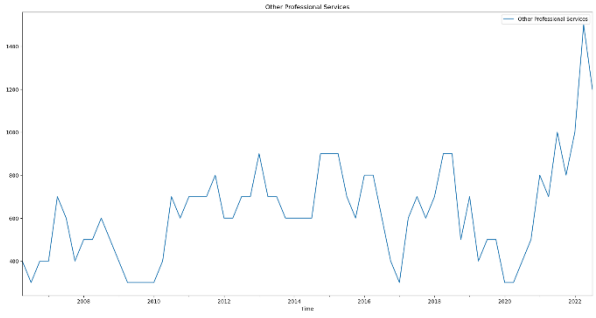




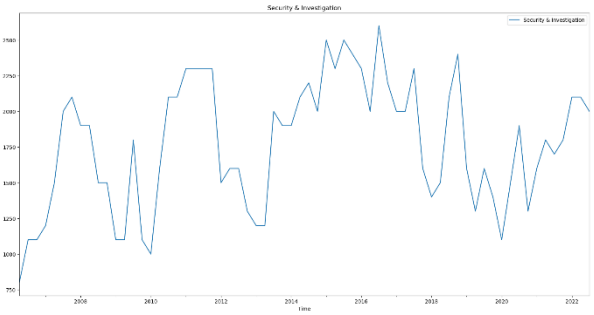


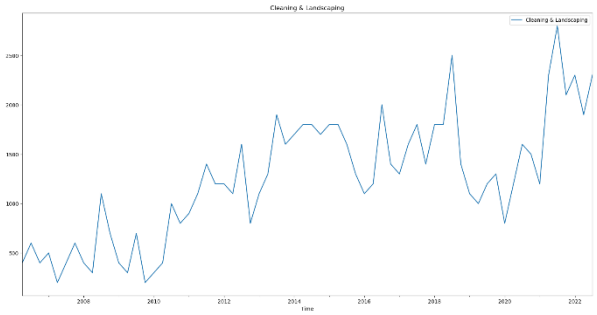


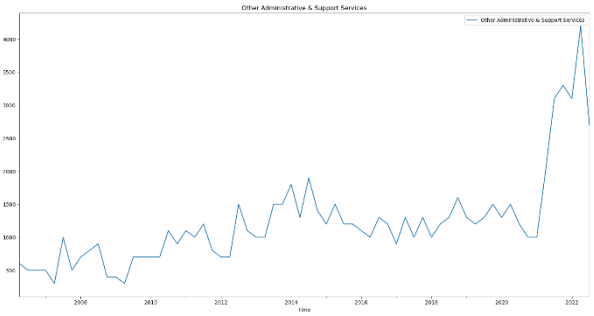


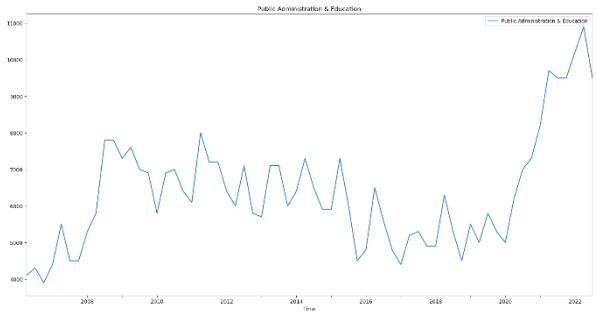


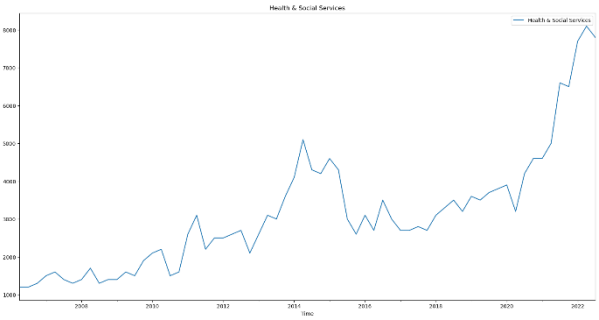
Click here to enter text.

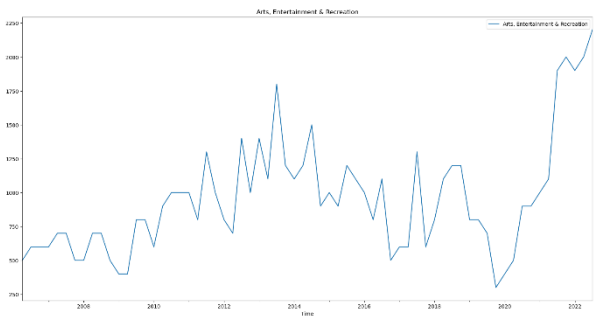


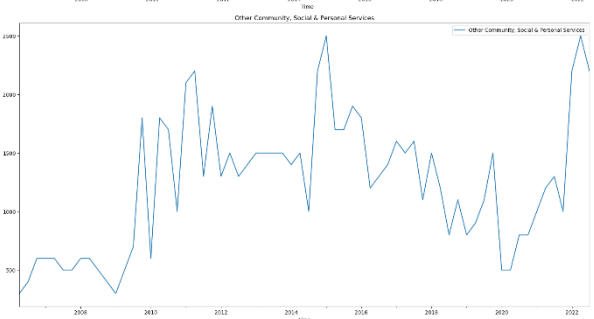


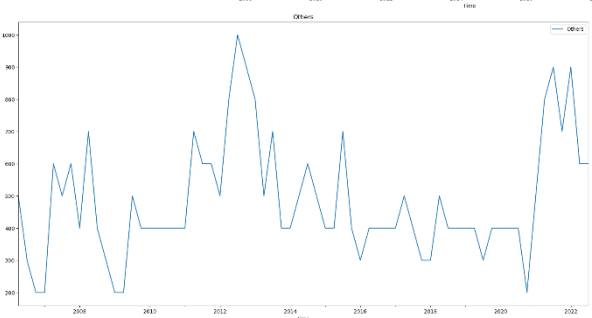










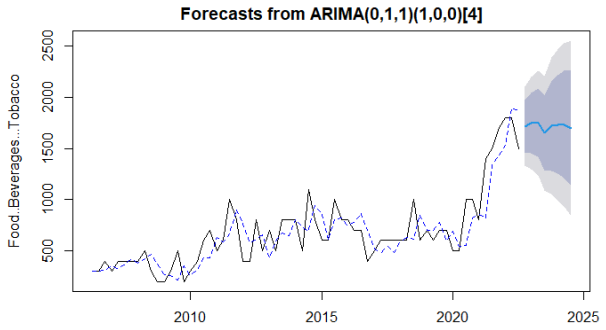


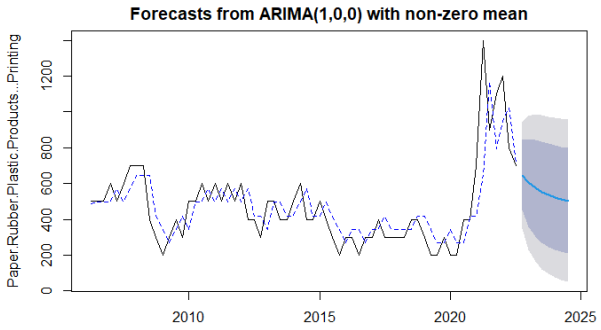
Click here to enter text.

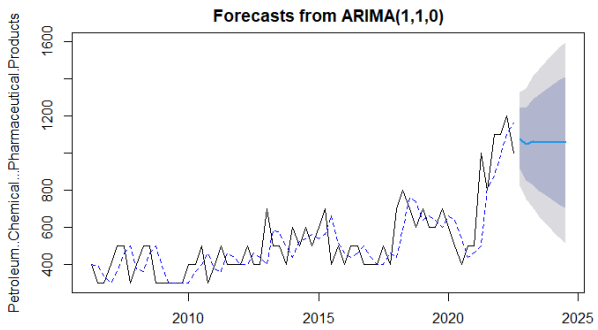
B3. ARIMA Model

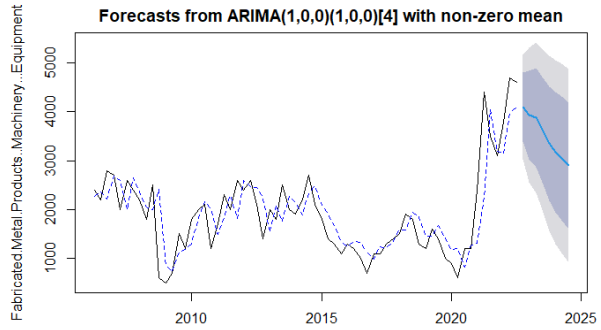
B3.1 ARIMA Forecast Plots

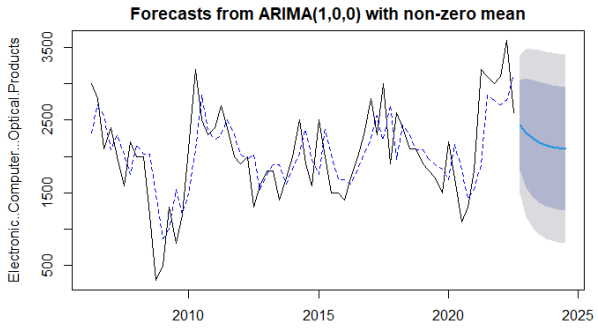
Below are the plots of the forecasts using ARIMA model. The auto-arima chosen model parameters are included in the title of each plot.

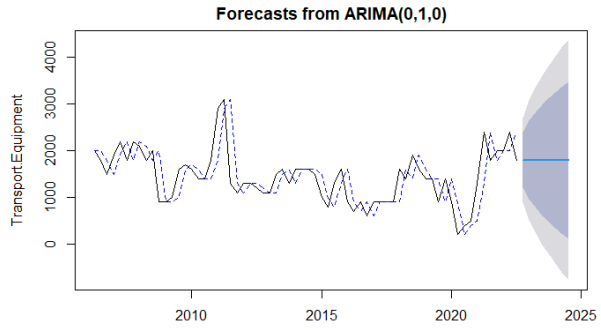




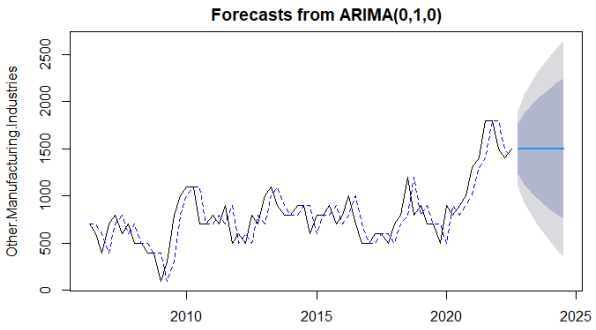


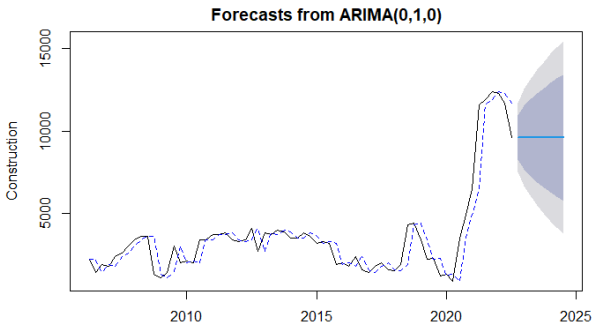


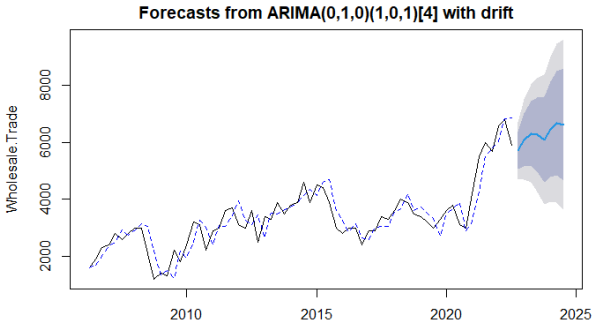


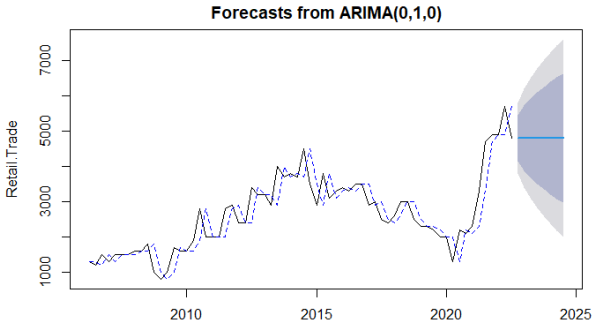


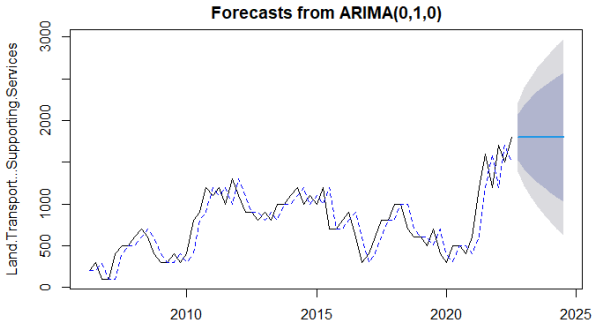
Click here to enter text.

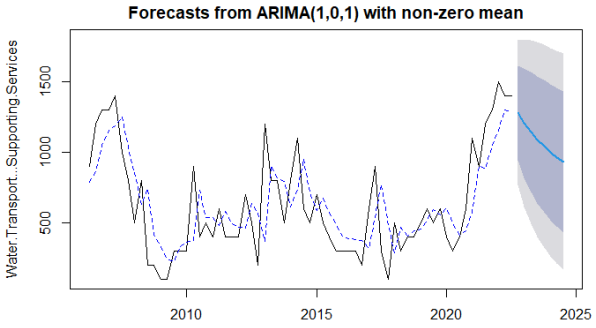


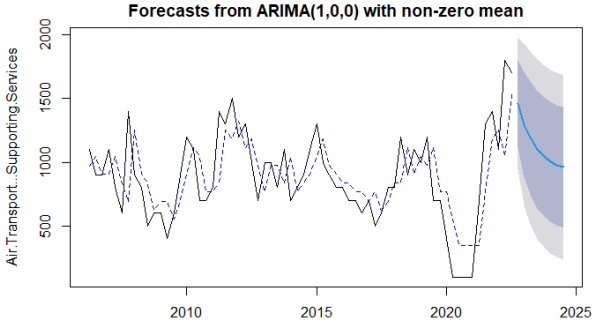


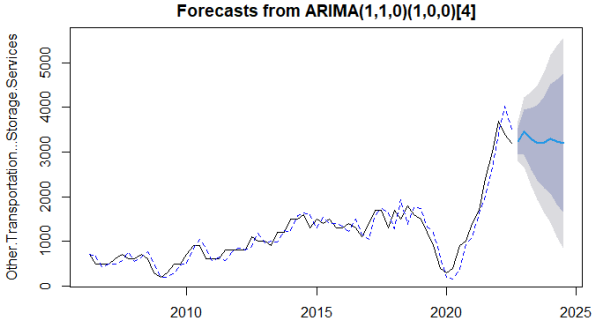




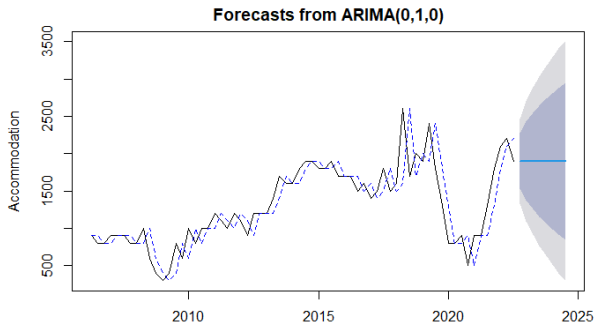


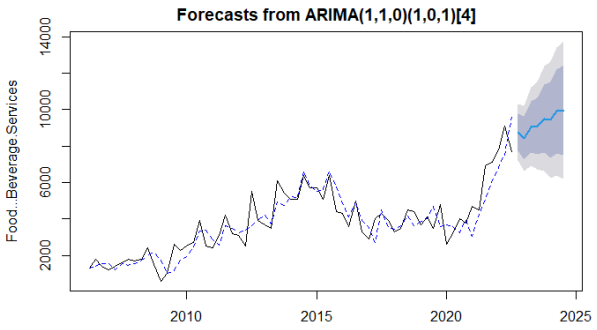


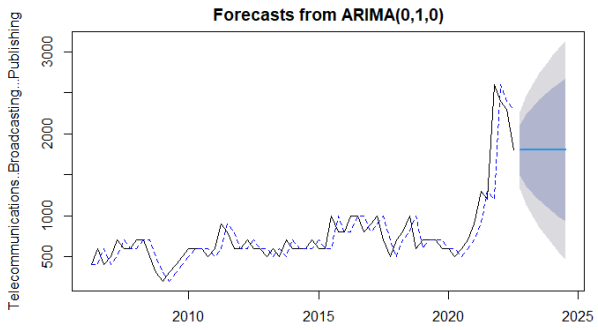


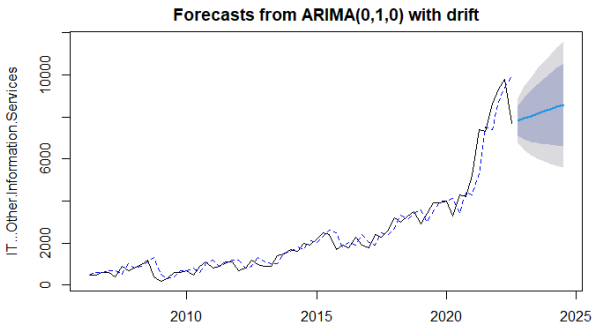


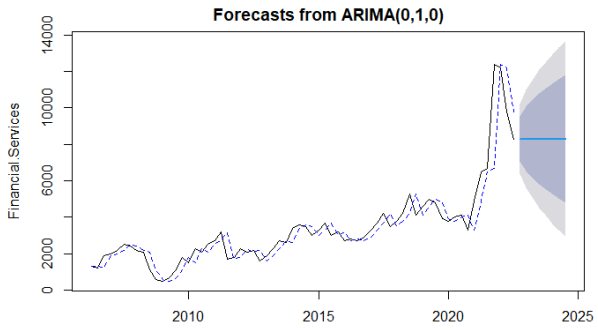
Click here to enter text.

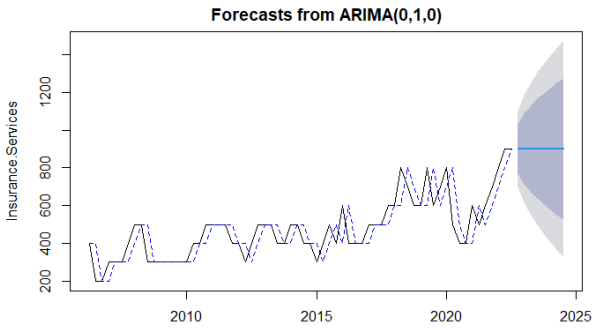


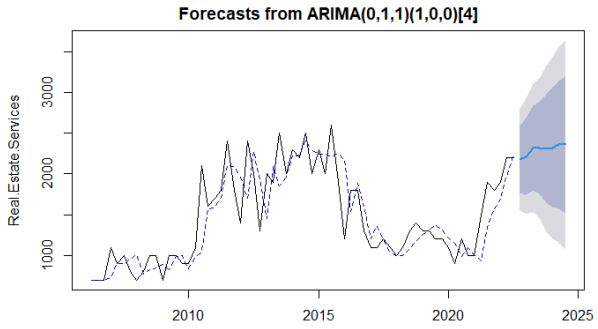


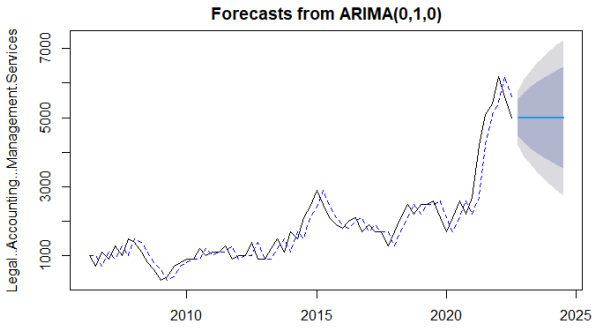




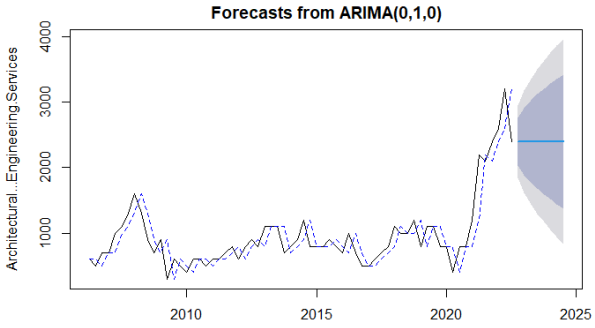


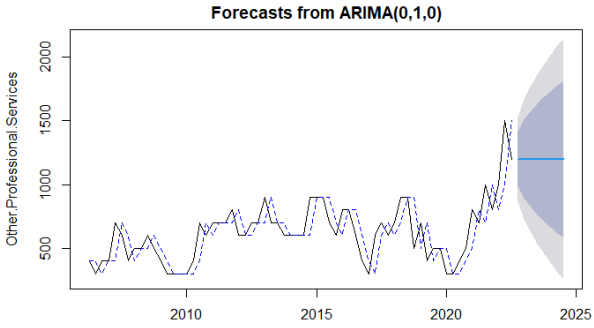


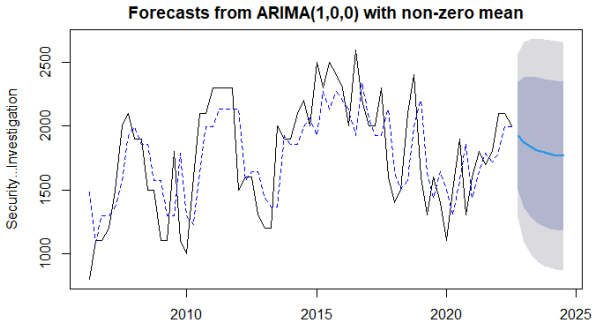


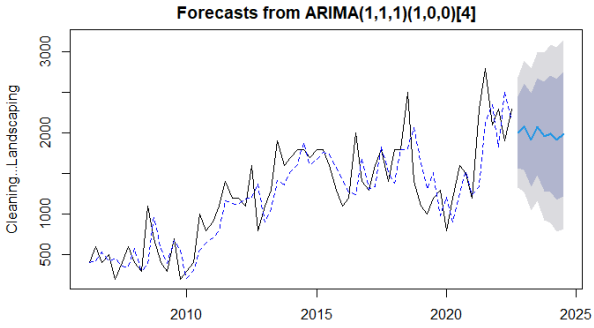


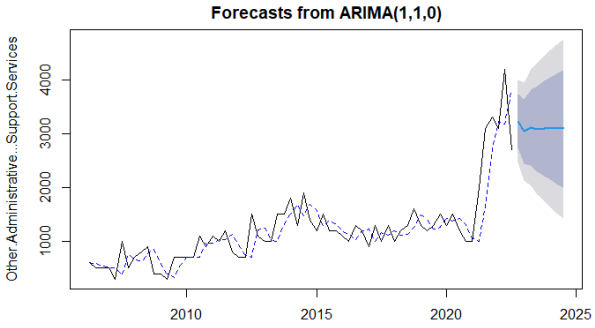
Click here to enter text.

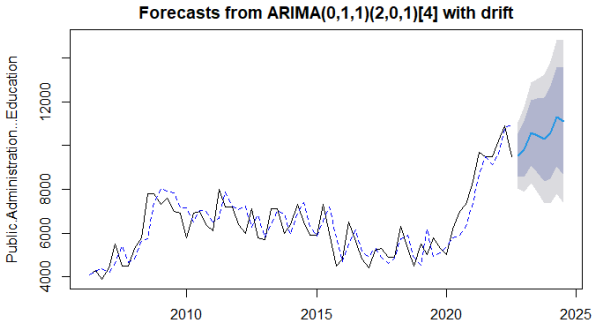




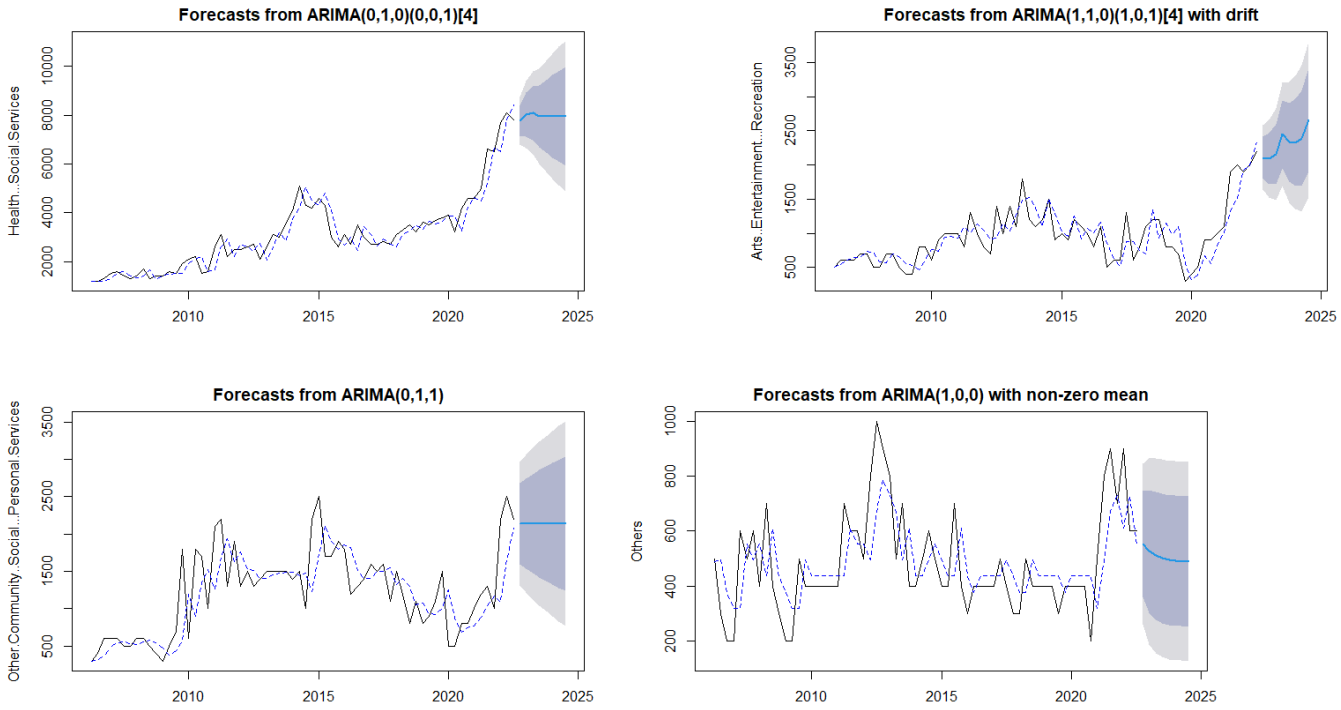








Click here to enter text.



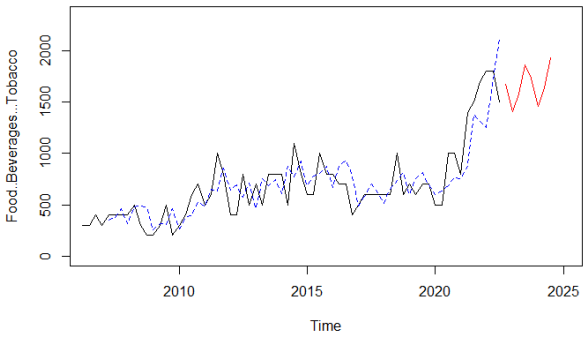
As shown in the plots, the auto arima algorithm chose the null model for many of the industries. This implies that the best predictor of future values is simply the current value. This is the case even when the time series shows a clear change in trend after 2020. As a result, the ARIMA model may not be suitable for our dataset, due to the changes in underlying factors and trends in the data over time.

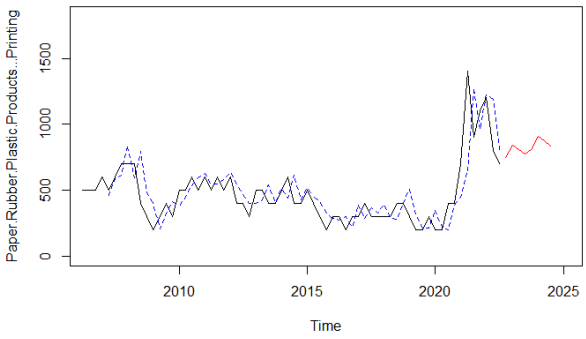
B4. Holt-Winters Model

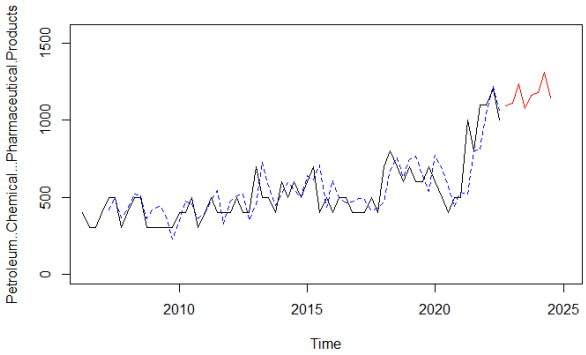
B4.1 Holt-Winters Forecast Plots

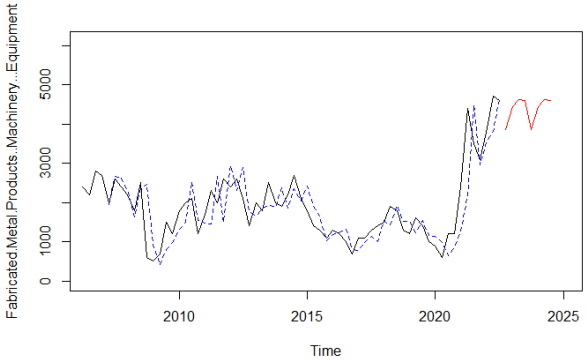
Below are the forecast plots for each industry using the Holt-Winters model.

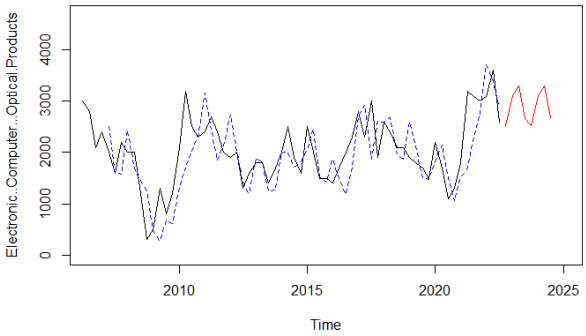
Click here to enter text.

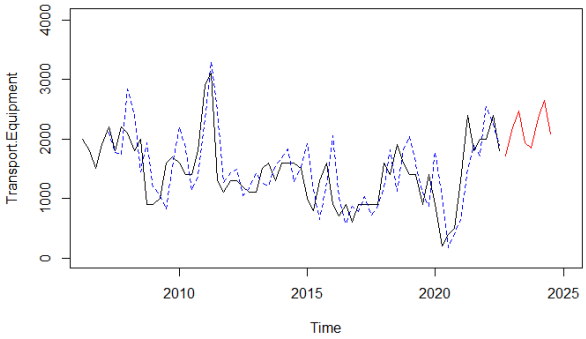


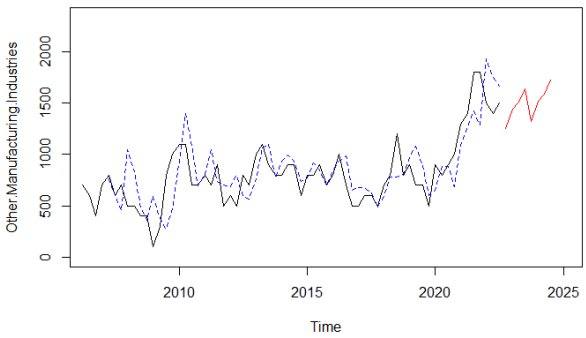


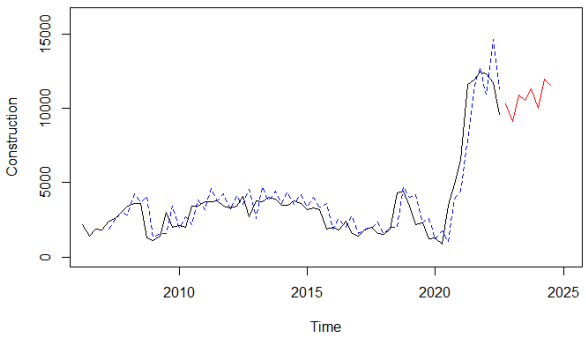




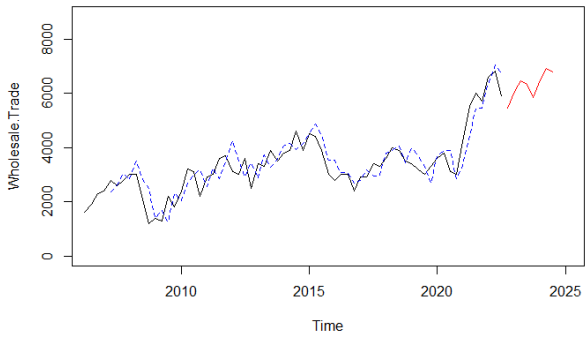


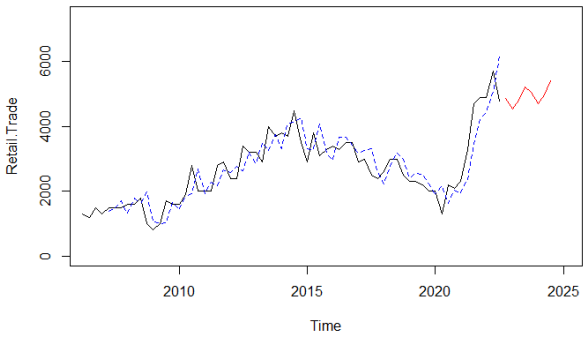


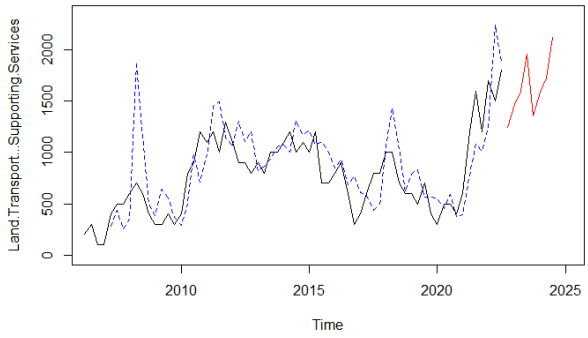


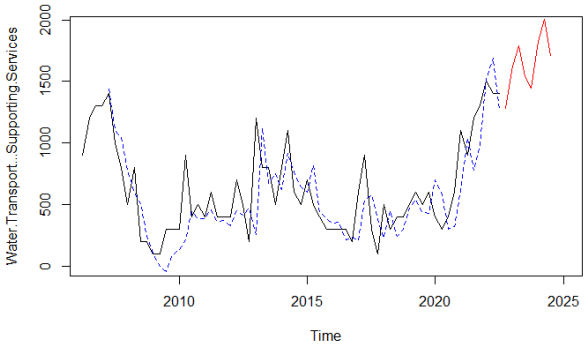


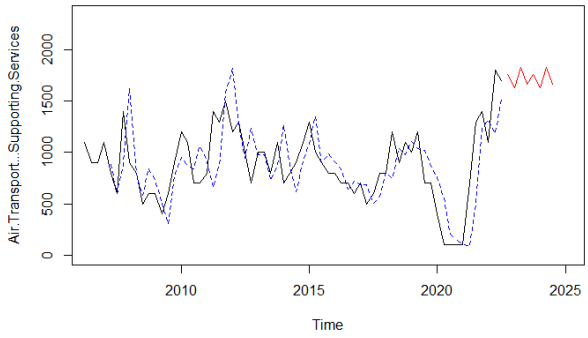
Click here to enter text.

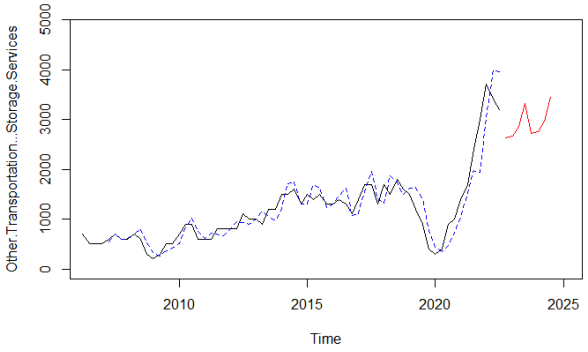


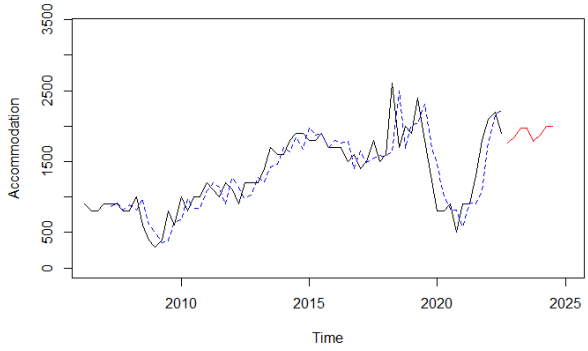


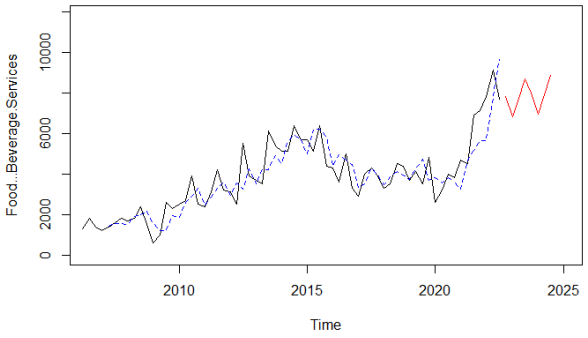




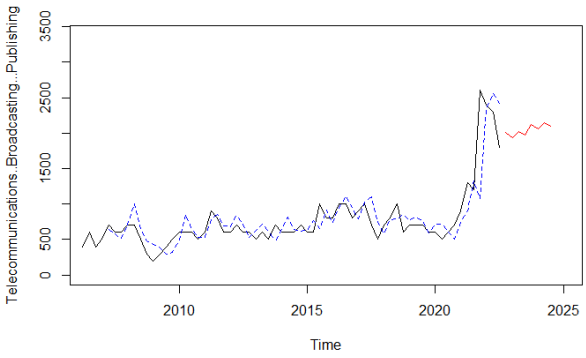


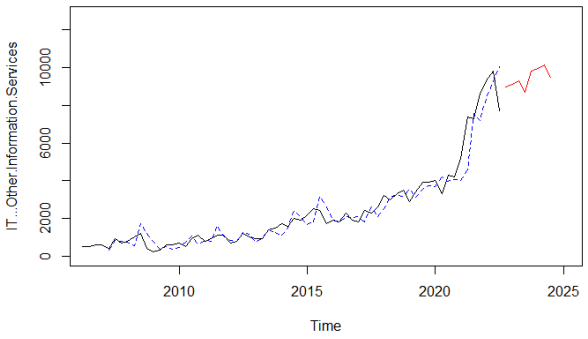


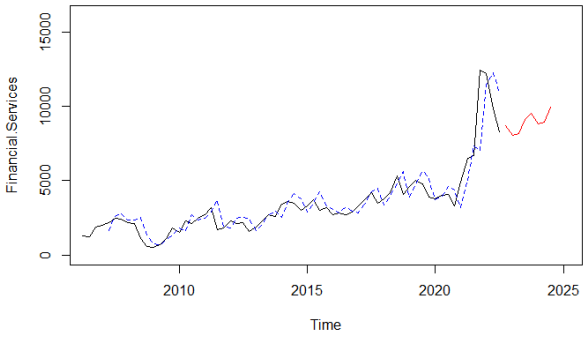


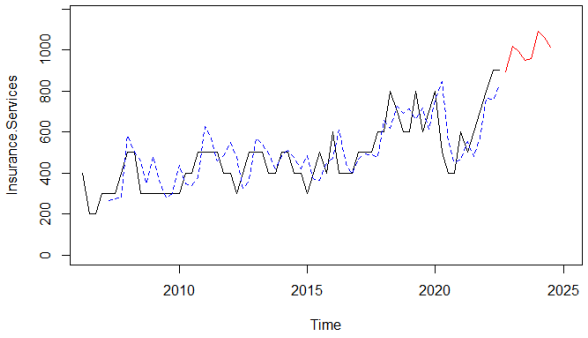


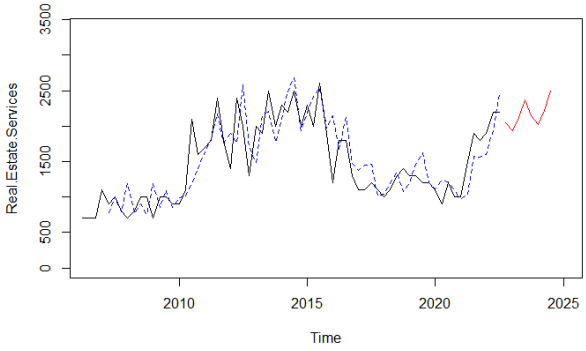
Click here to enter text.

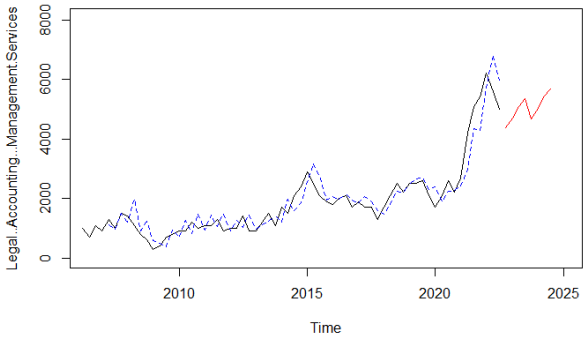


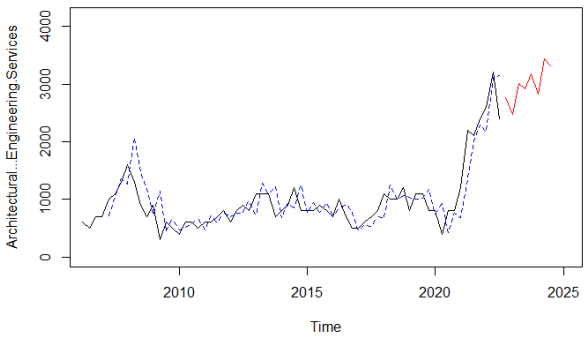


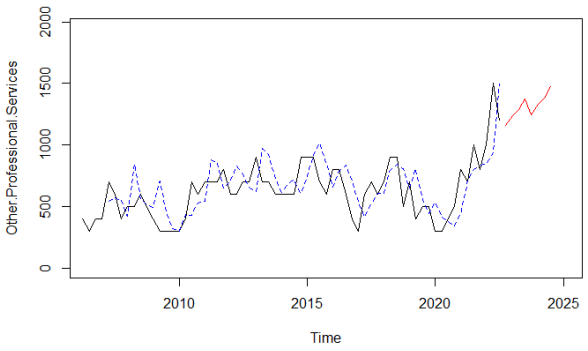




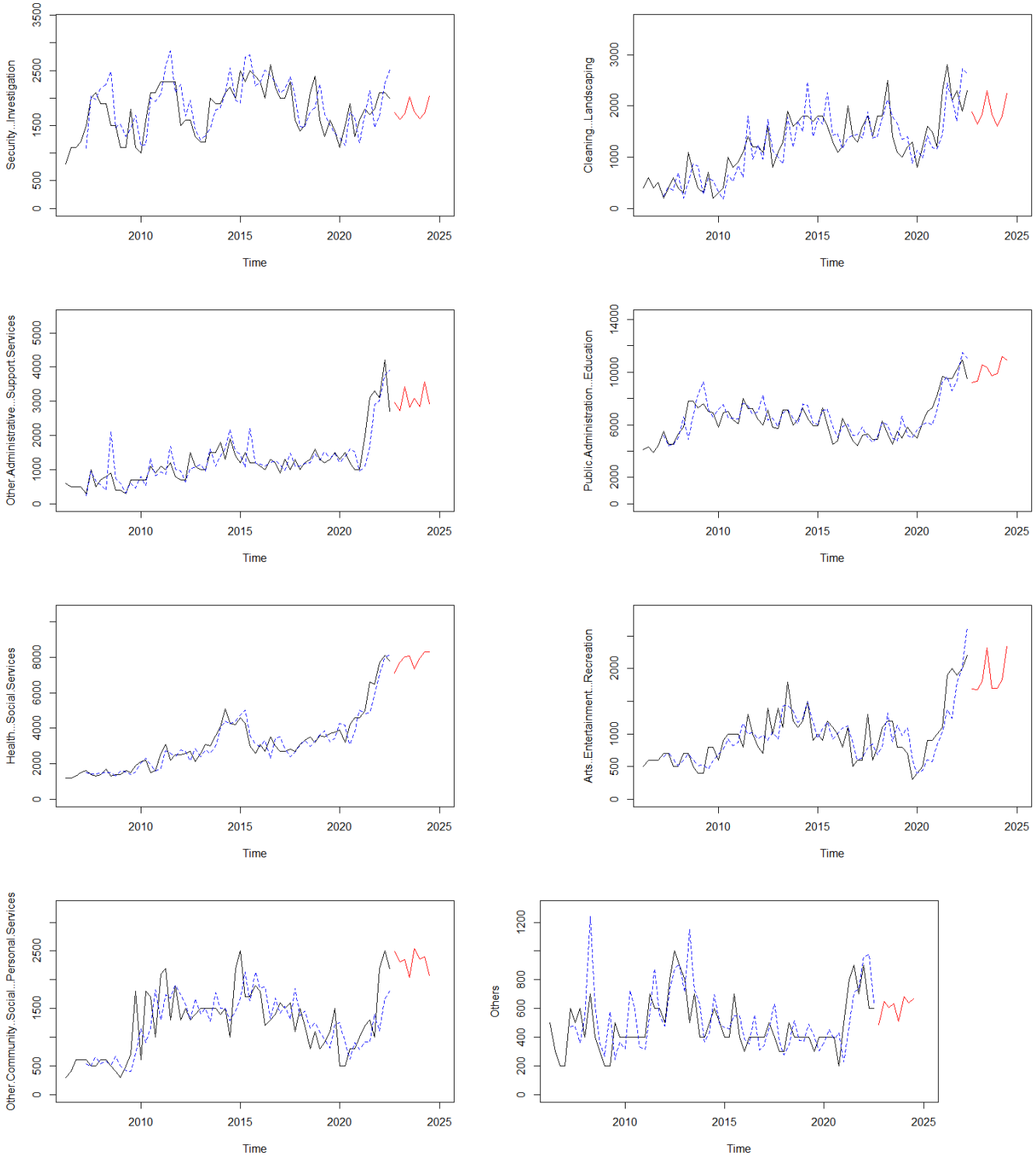








Click here to enter text.



While the Holt-Winters model fits better to the most recent trend, it seems to be overfitting the noise in its seasonal components. As a result, it tends to forecast seasonal trends that do not match the previous data points. The in-sample predictions also seems to deviate significantly from the actual values for some industries.

Click here to enter text.

B5. Taylor Expansion Model

B5.1 Model Parameters & Implementation

The Taylor Expansion Model estimates the derivatives of the time series data and uses the estimates as input to the Taylor series expansion to forecast future values.

Model Parameters

The model takes in 5 parameters:

Parameters	**Description**
Lag	Number of data points to average in the estimation of derivatives
Degree	The degree of approximation used (number of derivatives). E.g. degree = 1 means the model uses linear approximation, degree = 2 means that the model uses quadratic approximation
Freq	Frequency of time series data
Forecast Period	Number of periods to forecast out of sample
Time Step	Amount of time between each data point, used to estimate the derivatives accurately

Derivatives Calculation

The derivatives are estimated by differencing the time series data. For example, the first derivative at the current time is calculated by taking the current observation and subtracting the previous observation, divided by the time step.

$$y_i' = \frac{y_i - y_{i-1}}{\Delta t}$$

Where:

y_i': first derivative at index i

y_i: observation at index i

Δt: time step at each index

The second derivative is done by differencing the first derivative.

$$y_i^{(2)} = \frac{y_i' - y_{i-1}'}{\Delta t} = \frac{y_i - y_{i-1} - (y_{i-1} - y_{i-2})}{(\Delta t)^2} = \frac{y_i - 2y_{i-1} + y_{i-2}}{(\Delta t)^2}$$

Where:

$y_i^{(2)}$: second derivative at index i

y_i': first derivative at index i

y_i: observation at index i

Δt: time step at each index

This process can be done until the desired number of derivatives is estimated.

Derivatives Estimation

In order to be more robust to noise in the observations, a moving average of each derivative may be used instead of taking the latest value. This helps to smooth out the observations thus improving the estimate of the underlying trend in the data. For our model, a simple moving average is used. For example, the first and second derivatives are estimated by the following equations:

$$\widehat{y_i'} = \frac{1}{n}\sum_{k=0}^{n-1} y'_{i-k}$$

Where:

$\widehat{y_i'}$: first derivative estimate at index i

y'_i: first derivative at index i

n: smoothing parameter, how many periods to look back for calculation of moving average

$$\widehat{y_i^{(2)}} = \frac{1}{n}\sum_{k=0}^{n-1} y^{(2)}_{i-k}$$

Where:

$\widehat{y_i^{(2)}}$: second derivative estimate at index i

$y_i^{(2)}$: second derivative at index i

n: smoothing parameter, how many periods to look back for calculation of moving average

Forecasting

Once we have estimated the derivatives of our time series, we can forecast using Taylor series expansion.

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$$

Assume that we have data up until time t, and wish to forecast the time series at $t + \Delta t$. If we use degree of 1, we are basically extrapolating our data using linear approximation.

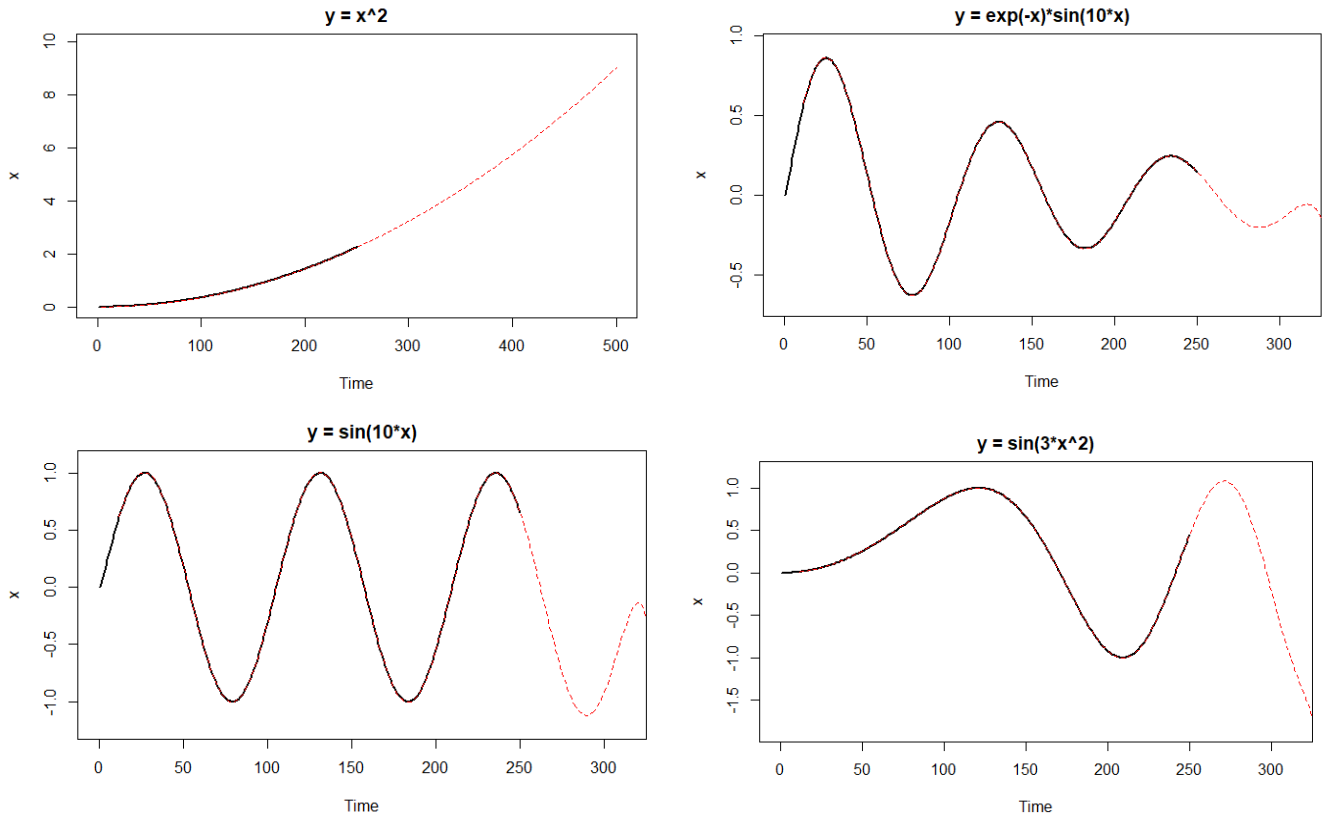
$$y_{t+\Delta t} = y_t + \frac{\widehat{y_i'}}{1!}(\Delta t)$$

If degree is set to 2, it is the same as quadratic approximation.

$$y_{t+\Delta t} = y_t + \frac{\widehat{y_i'}}{1!}(\Delta t) + \frac{\widehat{y_i^{(2)}}}{2!}(\Delta t)^2$$

As the number of degrees used increases, more terms are used from the Taylor series expansion formula to approximate future values. While this increases the precision of the forecasts, it also makes it more susceptible to noise, as higher order terms are taken to a higher power, thus amplifying any observation noise.

B5.2 Test Cases

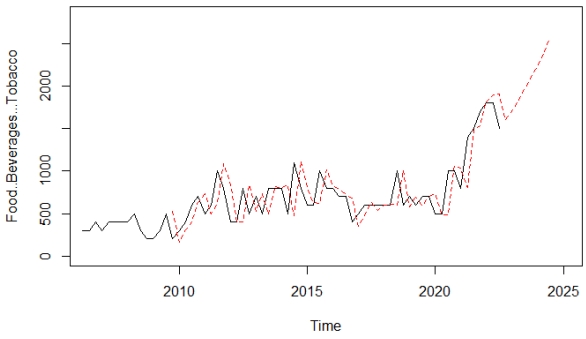


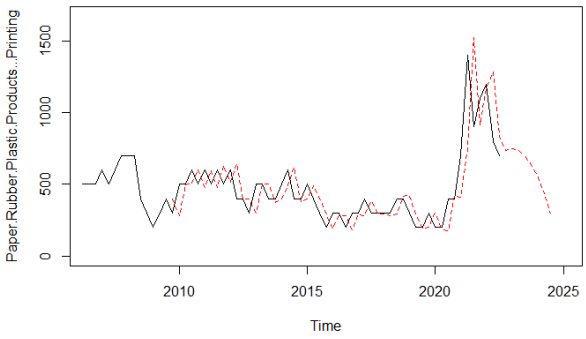
For the y = x^2 graph, degree is set to 2 as there a quadratic approximation is able to estimate the data perfectly. For the graphs with sinusoidal functions, we use degree = 12 to estimate the oscillation properties. In theory, there are an infinite amount of non-zero derivatives needed to describe the full graph.

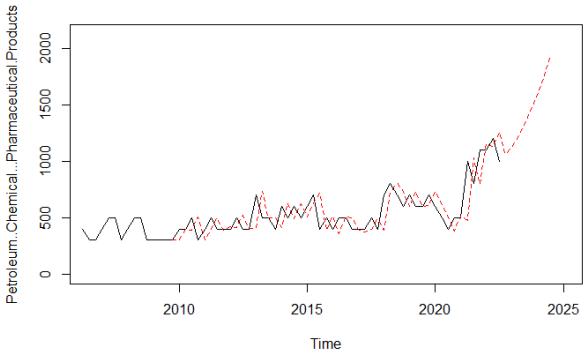
As the seasonality is not explicitly modelled, it is unable to forecast seasonal trends far into the future. However, when the degree used is high enough, it is still able to accurately model the curvature within the next cycle. The nonlinearity in the model also helps to implicitly measure changing magnitudes of the next cycle as well as changing frequencies. We can see that the model is able to forecast the damping characteristics in the second graph, and the increase in frequency in the fourth graph.

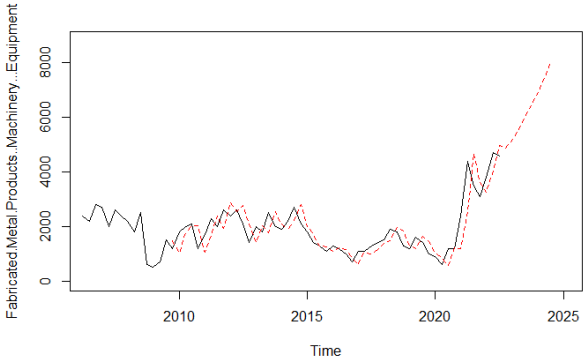
B5.3 Taylor Expansion Forecast Plots

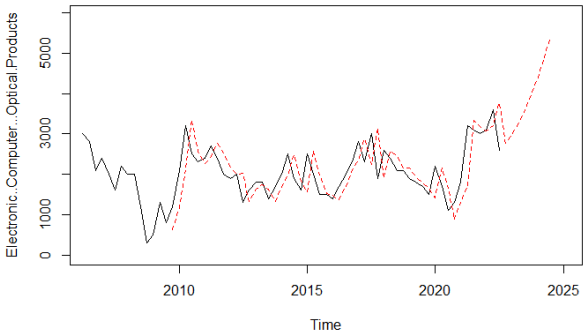
Below are the forecast plots for each industry using the Taylor Expansion model.

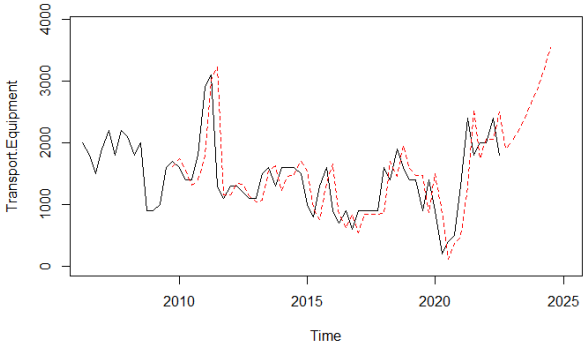


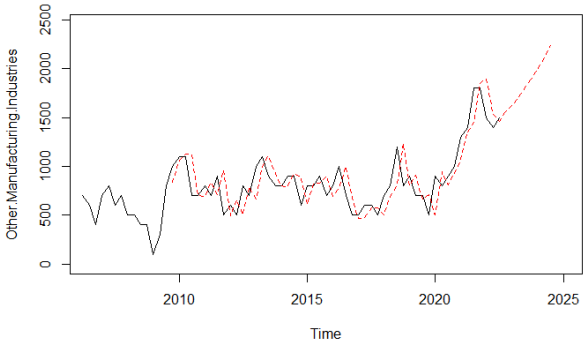


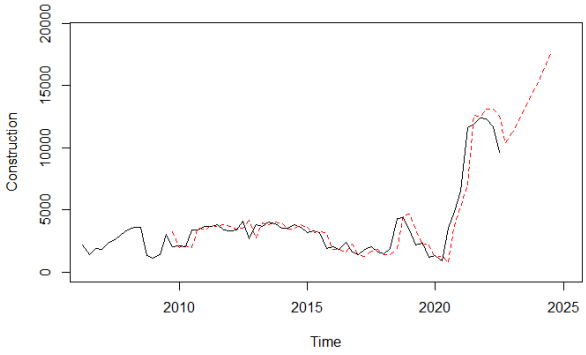




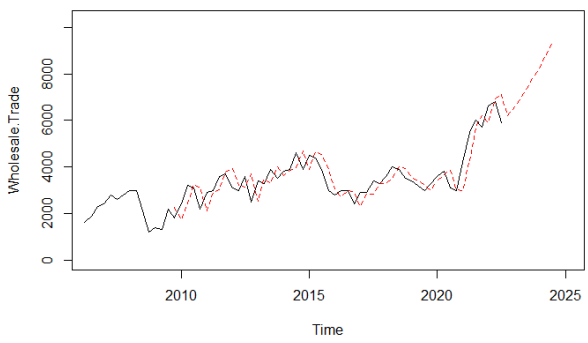


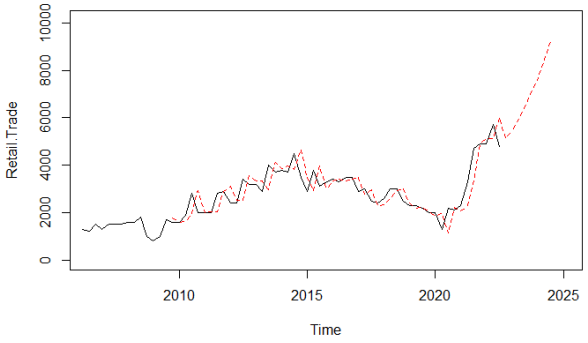


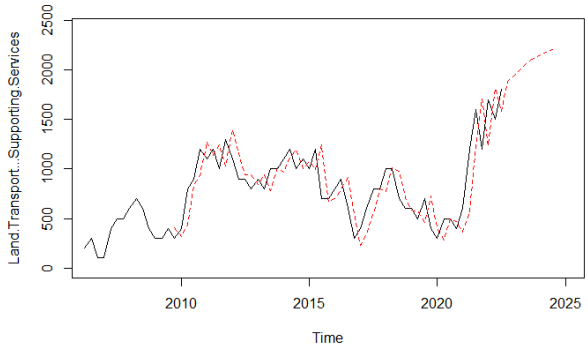


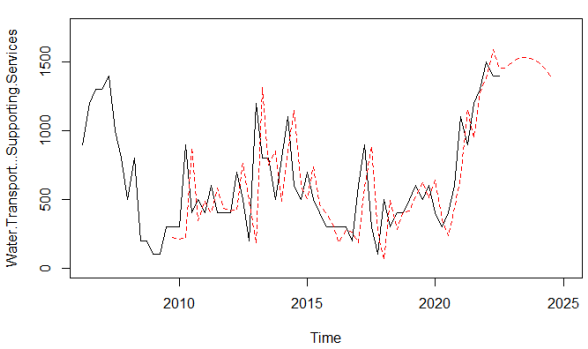


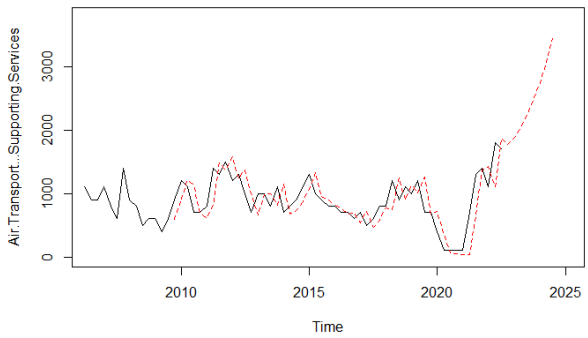
Click here to enter text.

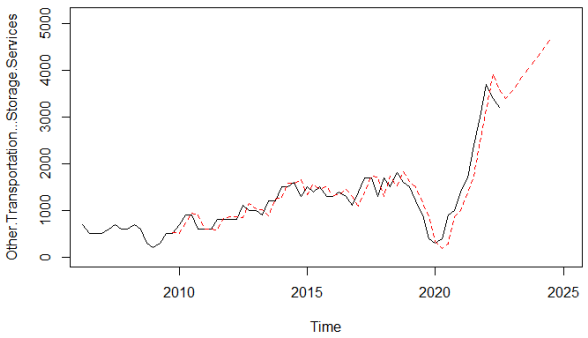


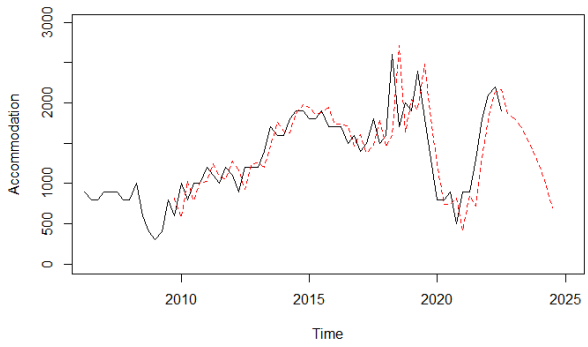


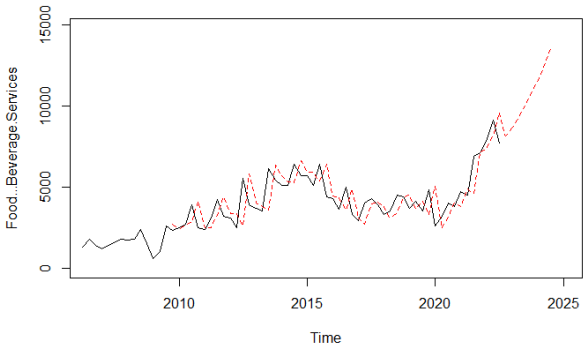




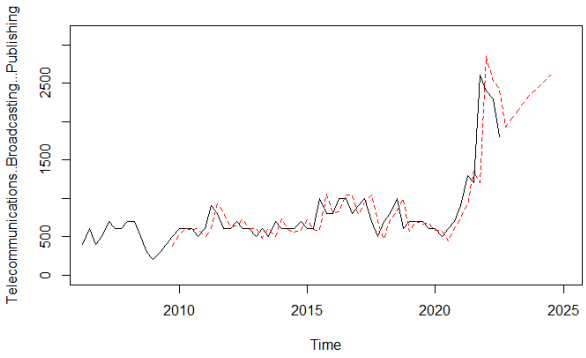


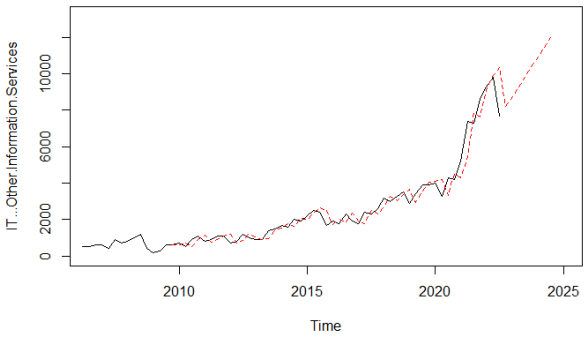


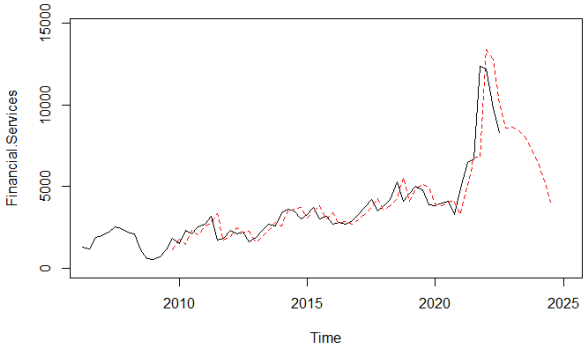


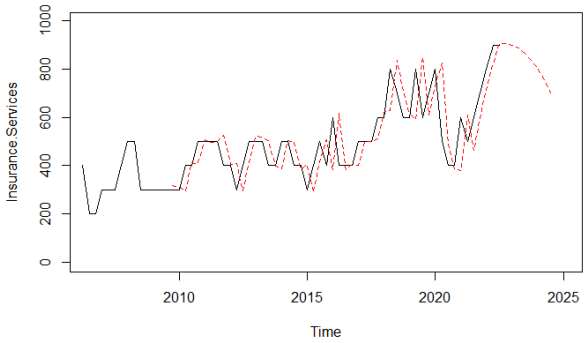


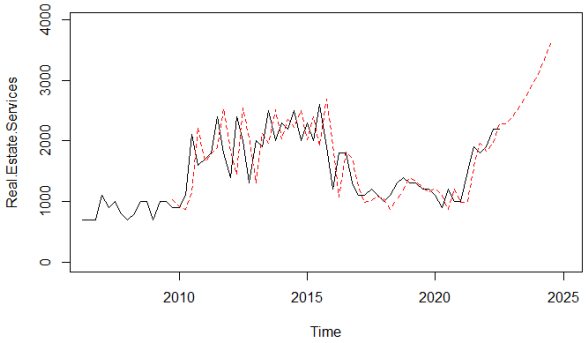
Click here to enter text.

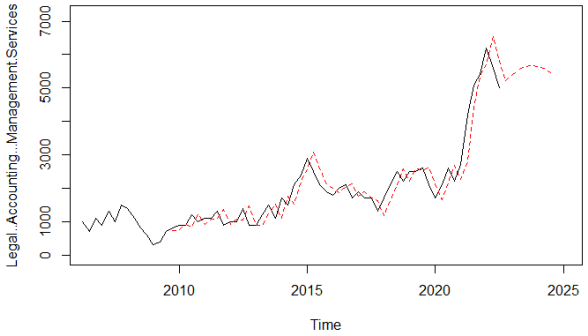


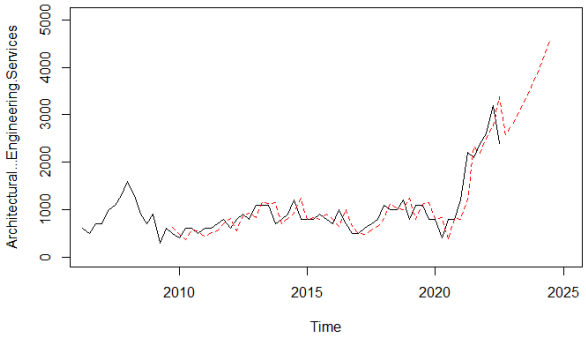


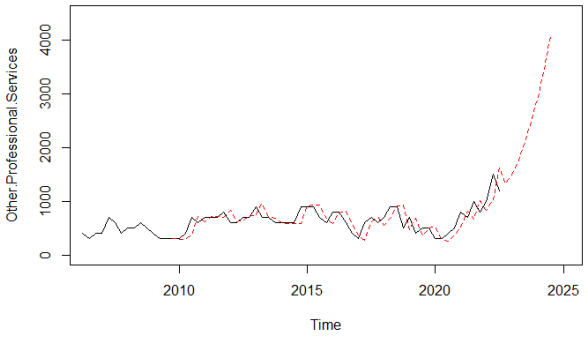




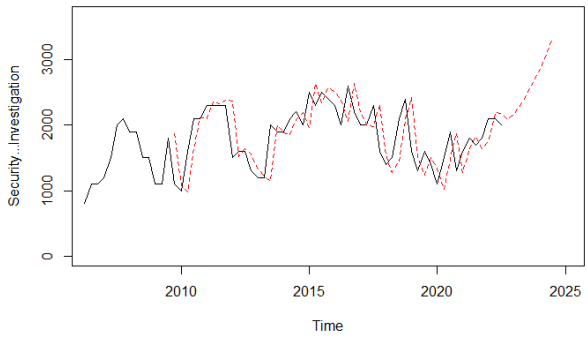


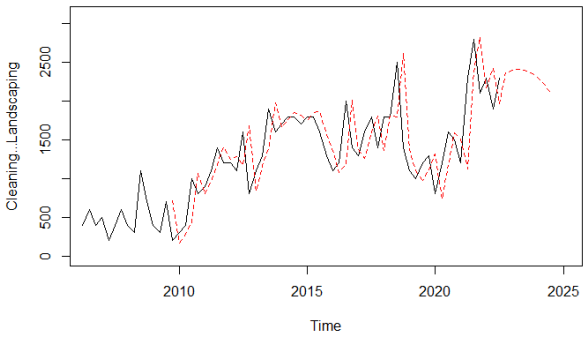


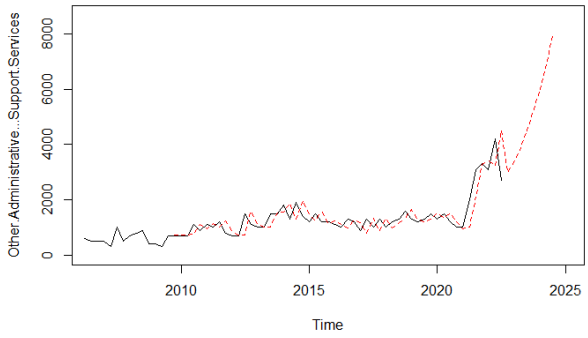


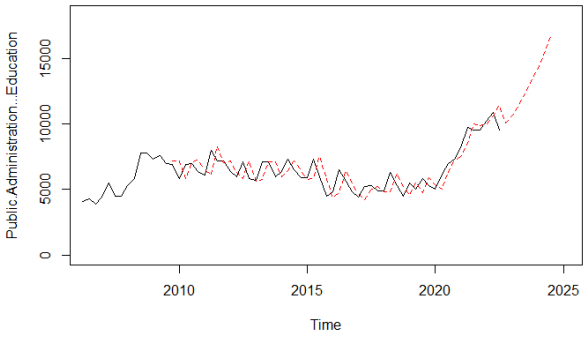


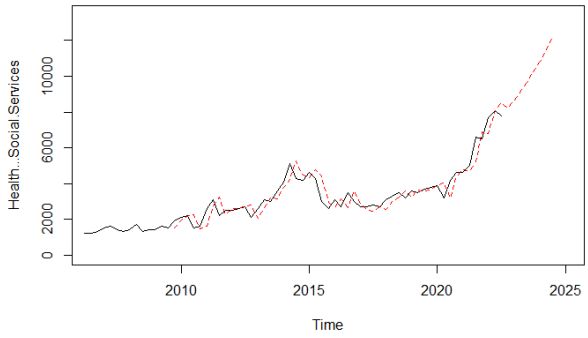
Click here to enter text.

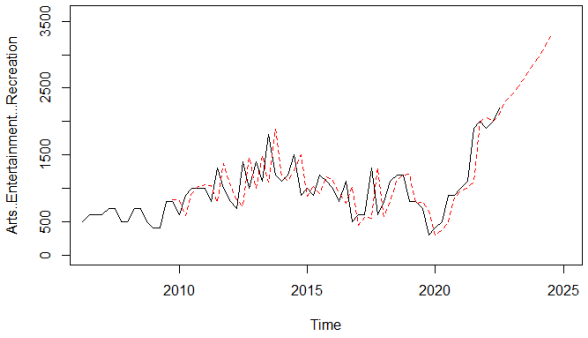


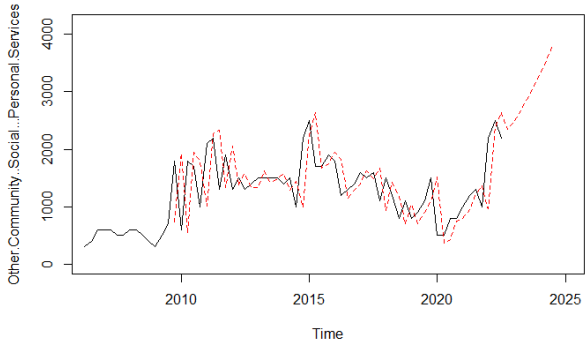


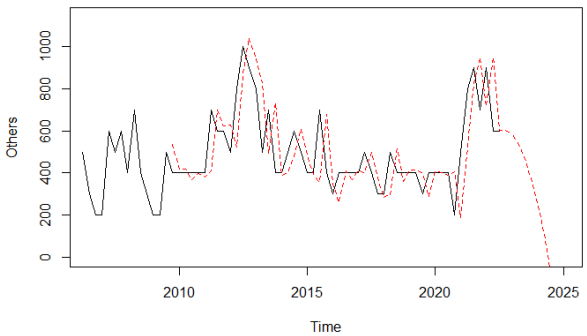












Click here to enter text.

B6. Mapping of Skills to Industry

By mapping the top skills in demand in each industry to our forecast of future industry demand, we can calculate an index of which skills are likely to be in demand in the future. To do this, we firstly need to forecast the future industry demand and estimate the weightage to assign each skill by each industry. We will represent these data as two separate matrices and then apply the following formula.

$$Industry\ Forecast\ Matrix \times Skill\ Weight\ Matrix = Skill\ Forecast\ Matrix$$

$$Industry\ Forecast\ Matrix = \begin{bmatrix} Q_1, Industry\ 1 & \cdots & Q_1, Industry\ n \\ \vdots & \ddots & \vdots \\ Q_p, Industry\ 1 & \cdots & Q_p, Industry\ n \end{bmatrix} \in \mathbb{R}^{p \times n}$$

The industry forecast matrix is a matrix of our demand forecast for each industry. It has p rows representing p quarters into the future and n columns representing the n different industries. In our case, p = 8 and n = 20. The forecast is estimated using the Taylor Expansion Model.

$$Skill\ Weight\ Matrix = \begin{bmatrix} Industry\ 1, Skill\ 1 & \cdots & Industry\ 1, Skill\ k \\ \vdots & \ddots & \vdots \\ Industry\ n, Skill\ 1 & \cdots & Industry\ n, Skill\ k \end{bmatrix} \in \mathbb{R}^{n \times k}$$

The skill weight matrix is a matrix of how important each skill is for a given industry. As it is a weightage, each row must add up to 1. It will have n rows representing the n different industries and k columns representing k different skills. In our case, n = 20 and k = 124.

$$Skill\ Forecast\ Matrix = \begin{bmatrix} Q_1, Skill\ 1 & \cdots & Q_1, Skill\ k \\ \vdots & \ddots & \vdots \\ Q_p, Skill\ 1 & \cdots & Q_p, Skill\ k \end{bmatrix} \in \mathbb{R}^{p \times k}$$

The skill forecast matrix is the output, calculated by multiplying the previously mentioned Industry Forecast Matrix with the Skill Weight Matrix. It will have p rows representing p quarters forecasted into the future and k columns representing k different skills. In our case, p = 8 and k = 124.

Putting it all together:

$$\begin{aligned} &Industry\ Forecast\ Matrix \times Skill\ Weight\ Matrix \\ &= \begin{bmatrix} Q_1, Industry\ 1 & \cdots & Q_1, Industry\ n \\ \vdots & \ddots & \vdots \\ Q_p, Industry\ 1 & \cdots & Q_p, Industry\ n \end{bmatrix} \begin{bmatrix} Industry\ 1, Skill\ 1 & \cdots & Industry\ 1, Skill\ k \\ \vdots & \ddots & \vdots \\ Industry\ n, Skill\ 1 & \cdots & Industry\ n, Skill\ k \end{bmatrix} \\ &= \begin{bmatrix} Q_1, Skill\ 1 & \cdots & Q_1, Skill\ k \\ \vdots & \ddots & \vdots \\ Q_p, Skill\ 1 & \cdots & Q_p, Skill\ k \end{bmatrix} = Skill\ Forecast\ Matrix \end{aligned}$$

B6.1 Skill Weight Matrix

The original industry skills dataset consists of the top 10 skills in demand from each industry. This data is retrieved using text mining techniques, by looking at the most common keywords used in job postings in different industries. Below is a sample from the dataset.

	industry_name	skill_group_name	skill_group_rank
10	Oil & Energy	Oil & Gas	1.0
11	Oil & Energy	Drilling Engineering	2.0
12	Oil & Energy	Utilities	3.0
13	Oil & Energy	Negotiation	4.0
14	Oil & Energy	Digital Literacy	5.0
...	...	...	...
685	Health, Wellness & Fitness	Sports Coaching	6.0
686	Health, Wellness & Fitness	Digital Literacy	7.0
687	Health, Wellness & Fitness	Public Health	8.0
688	Health, Wellness & Fitness	Business Management	9.0
689	Health, Wellness & Fitness	Teamwork	10.0

430 rows × 3 columns

However, as the industry classification in this dataset is different from the industry demand dataset, we need to match the industries between the two dataset. This is done manually by going through the industries in the industry skills dataset and categorising each one under one of the original industries.

industry forecast dataset	industry skills dataset							
Food, Beverages & Tobacco	Food Production	Packaging & Containers						
Paper/Rubber/Plastic Products & Printing	Paper & Forest Products	Printing	Plastics					
Petroleum, Chemical & Pharmaceutical Products	Oil & Energy	Pharmaceuticals	Chemicals					
Fabricated Metal Products, Machinery & Equipment	Machinery	Industrial Automation						
Electronic, Computer & Optical Products	Electrical & Electronic Manufacturing	Computer Hardware	Semiconductors					
Transport Equipment								
Other Manufacturing Industries	Mechanical Or Industrial Engineering							
Construction								
Wholesale Trade								
Retail Trade								
Land Transport & Supporting Services	Automotive							
Water Transport & Supporting Services								
Air Transport & Supporting Services	Aviation & Aerospace							
Other Transportation & Storage Services								
Accommodation								
Food & Beverage Services								
Telecommunications, Broadcasting & Publishing	Computer Networking	Internet	Telecommunications					
IT & Other Information Services	Computer Software	Information Technology & Services	Computer & Netw	Information Services				
Financial Services	Banking	Financial Servic	Investment Banking	Investment Mana	Venture Capital & Private Equity			
Insurance Services	Insurance							
Real Estate Services								
Legal, Accounting & Management Services	Law Practice	Legal Services	Management Consulting	Accounting	Outsourcing/Offshoring			
Architectural & Engineering Services	Architecture & Planning							
Other Professional Services								
Security & Investigation								
Cleaning & Landscaping								
Other Administrative & Support Services	Marketing & Advertising	Executive Office	Market Research	Public Relations & Communications				
Public Administration & Education	Environmental Services							
Health & Social Services	Veterinary	Health, Wellness & Fitness						
Arts, Entertainment & Recreation	Motion Pictures & Film	Broadcast Media	Newspapers	Publishing	Writing & Editing	Computer Games	Online Media	Media P
Other Community, Social & Personal Services	Events Services							
Others	Mining & Metals	Textiles	Renewables & Environment	Biotechnology	Research	Design	Professional Training	

Click here to enter text.

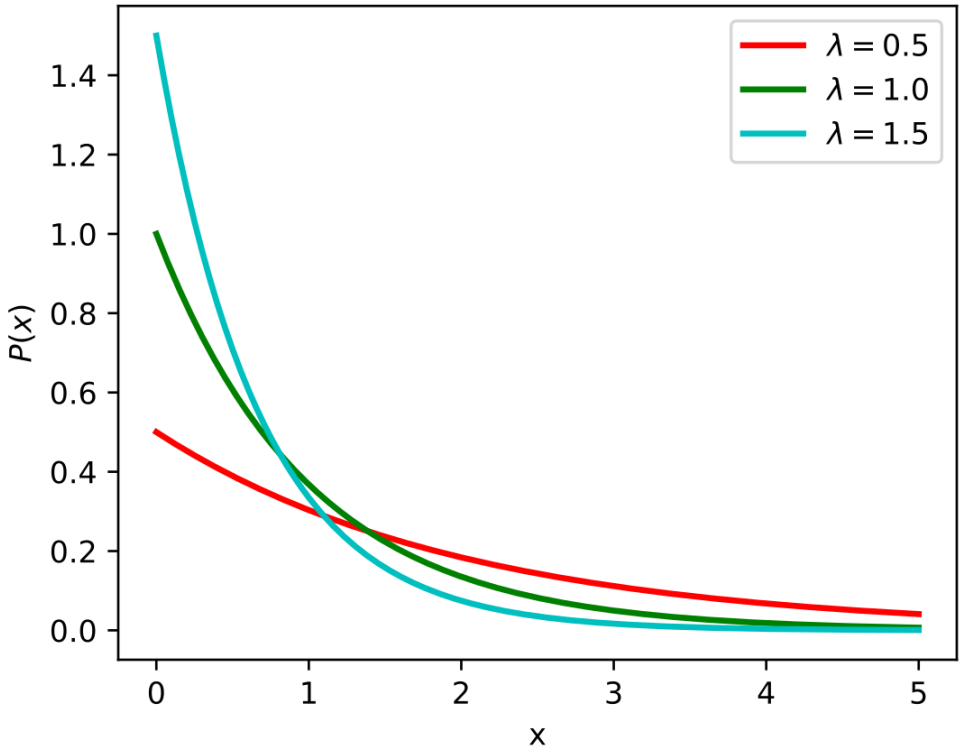
Some industries are mapped to multiple industries in the corresponding dataset, while some do not have any mapping. As such, we remove those without any mapping, and keep the top 3 most appropriate ones for those with multiple mappings.

industry forecast dataset	industry skills dataset		
	industry 1	industry 2	industry 3
Food, Beverages & Tobacco	Food Production	Packaging & Containers	
Paper/Rubber/Plastic Products & Printing	Paper & Forest Products	Printing	Plastics
Petroleum, Chemical & Pharmaceutical Products	Oil & Energy	Pharmaceuticals	Chemicals
Fabricated Metal Products, Machinery & Equipment	Machinery	Industrial Automation	
Electronic, Computer & Optical Products	Electrical & Electronic Manufacturing	Computer Hardware	Semiconductors
Other Manufacturing Industries	Mechanical Or Industrial Engineering		
Land Transport & Supporting Services	Automotive		
Air Transport & Supporting Services	Aviation & Aerospace		
Telecommunications, Broadcasting & Publishing	Computer Networking	Internet	Telecommunications
IT & Other Information Services	Computer Software	Information Technology & Services	Information Services
Financial Services	Banking	Financial Services	Investment Banking
Insurance Services	Insurance		
Legal, Accounting & Management Services	Accounting	Legal Services	Management Consulting
Architectural & Engineering Services	Architecture & Planning		
Other Administrative & Support Services	Marketing & Advertising	Public Relations & Communications	Market Research
Public Administration & Education	Environmental Services		
Health & Social Services	Veterinary	Health, Wellness & Fitness	
Arts, Entertainment & Recreation	Entertainment	Performing Arts	Media Production
Other Community, Social & Personal Services	Events Services		
Others	Research	Biotechnology	Renewables & Environment

This leaves us with just 20 industries from the industry demand forecast dataset. The skill ranking for each industry is then recalculated by taking the average of the rankings from the mapped industries in the skills dataset. A matrix is then created, with the rows being the different industries and the columns being the different skills. Each entry is a number between 0 to 10, showing the ranking of each skill for each industry.

	Oil & Gas	Drilling Engineering	Utilities	Negotiation	Digital Literacy	Project Management	Teamwork	Leadership	Business Management	Instrumentation	...	Event Planning
Food, Beverages & Tobacco	0.0	0.0	0.0	3.333333	4.000000	0.000000	4.000000	3.333333	4.333333	0.0	...	0.000000
Paper/Rubber/Plastic Products & Printing	0.0	0.0	0.0	3.500000	3.500000	4.500000	5.000000	6.250000	4.750000	0.0	...	0.000000
Petroleum, Chemical & Pharmaceutical Products	0.5	1.0	1.5	2.333333	4.250000	5.333333	4.500000	6.000000	6.750000	5.0	...	0.000000
Fabricated Metal Products, Machinery & Equipment	0.0	1.5	2.5	0.500000	4.333333	5.000000	3.500000	2.500000	4.500000	2.0	...	0.000000
Electronic, Computer & Optical Products	0.0	4.0	5.0	0.000000	2.750000	4.500000	2.000000	4.000000	3.500000	0.0	...	0.000000
Other Manufacturing Industries	0.0	2.5	0.0	3.000000	1.500000	4.000000	5.000000	4.500000	0.000000	0.0	...	0.000000
Land Transport & Supporting Services	0.0	0.0	0.0	1.000000	2.500000	0.000000	3.000000	1.500000	3.500000	0.0	...	0.000000
Air Transport & Supporting Services	0.0	0.0	0.0	0.000000	3.500000	0.000000	4.000000	5.000000	0.000000	0.0	...	0.000000
Telecommunications, Broadcasting & Publishing	0.0	0.0	0.0	0.000000	5.250000	3.500000	2.333333	5.000000	3.000000	0.0	...	0.000000
IT & Other Information Services	0.0	0.0	0.0	0.000000	1.000000	6.000000	4.000000	6.250000	1.500000	0.0	...	0.000000
Financial Services	0.0	0.0	0.0	0.000000	5.750000	0.000000	3.500000	4.500000	6.000000	0.0	...	0.000000
Insurance Services	0.0	0.0	0.0	3.000000	1.500000	0.000000	1.000000	2.000000	2.500000	0.0	...	0.000000
Legal, Accounting & Management Services	0.0	0.0	0.0	5.333333	6.250000	2.000000	4.666667	1.000000	0.500000	0.0	...	0.000000
Architectural & Engineering Services	0.0	0.0	0.0	0.000000	4.000000	3.000000	0.000000	0.000000	0.000000	0.0	...	0.000000

Next, we need to transform these rankings into a weightage. As a lower number in ranking means a higher importance, we transform this data by passing the ranking into the exponential distribution with parameter lambda = 5 while ignoring entries that are 0.



The exponential distribution was chosen as a heuristic due to its downward slope, which gives more weight to lower ranks and less weight to higher ranks. The resulting matrix is then normalised to ensure that each row adds up to 1.

Thus, our final Skills Weight Matrix looks something like this, with 20 rows and 124 columns:

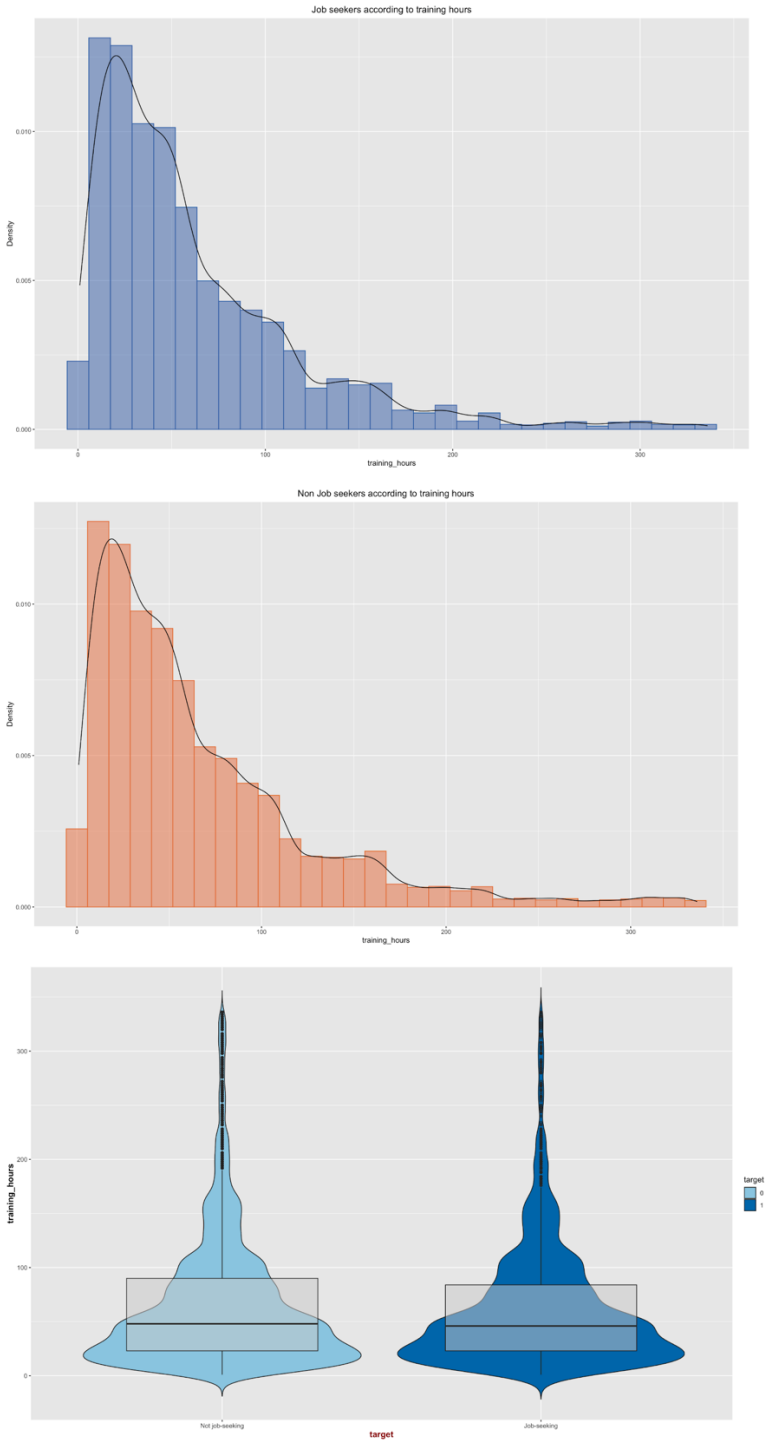
	Oil & Gas	Drilling Engineering	Utilities	Negotiation	Digital Literacy	Project Management	Teamwork	Leadership	Business Management	Instrumentation	...	Event Planning
Food, Beverages & Tobacco	0.000000	0.000000	0.000000	0.037295	0.011233	0.000000	0.011233	0.037295	0.004132	0.000000	...	0.000000
Paper/Rubber/Plastic Products & Printing	0.000000	0.000000	0.000000	0.014947	0.014947	0.004574	0.002077	0.001227	0.003515	0.000000	...	0.000000
Petroleum, Chemical & Pharmaceutical Products	0.201445	0.092228	0.057715	0.022601	0.012098	0.002536	0.005539	0.001855	0.001357	0.003466	...	0.000000
Fabricated Metal Products, Machinery & Equipment	0.000000	0.106156	0.034325	0.237785	0.004217	0.001602	0.013042	0.034325	0.003054	0.065434	...	0.000000
Electronic, Computer & Optical Products	0.000000	0.004905	0.001999	0.000000	0.033561	0.002937	0.082335	0.004905	0.013680	0.000000	...	0.000000
Other Manufacturing Industries	0.000000	0.046371	0.000000	0.026605	0.140862	0.008758	0.002883	0.005025	0.000000	0.000000	...	0.000000
Land Transport & Supporting Services	0.000000	0.000000	0.000000	0.245510	0.046371	0.000000	0.026605	0.140862	0.015265	0.000000	...	0.000000
Air Transport & Supporting Services	0.000000	0.000000	0.000000	0.000000	0.015265	0.000000	0.008758	0.002883	0.000000	0.000000	...	0.000000
Telecommunications, Broadcasting & Publishing	0.000000	0.000000	0.000000	0.000000	0.001262	0.015379	0.061677	0.002200	0.020304	0.000000	...	0.000000
IT & Other Information Services	0.000000	0.000000	0.000000	0.000000	0.249373	0.002372	0.007930	0.001680	0.148667	0.000000	...	0.000000
Financial Services	0.000000	0.000000	0.000000	0.000000	0.005338	0.000000	0.042872	0.015128	0.003519	0.000000	...	0.000000
Insurance Services	0.000000	0.000000	0.000000	0.026605	0.140862	0.000000	0.245510	0.080820	0.046371	0.000000	...	0.000000
Legal, Accounting & Management Services	0.000000	0.000000	0.000000	0.001526	0.001222	0.030657	0.002973	0.104071	0.181387	0.000000	...	0.000000
Architectural & Engineering Services	0.000000	0.000000	0.000000	0.000000	0.008758	0.026605	0.000000	0.000000	0.000000	0.000000	...	0.000000
Other Administrative & Support Services	0.000000	0.000000	0.000000	0.000000	0.002059	0.006729	0.004534	0.001582	0.000000	0.000000	...	0.081972

Click here to enter text.

Appendix C –Job Seeker Prediction

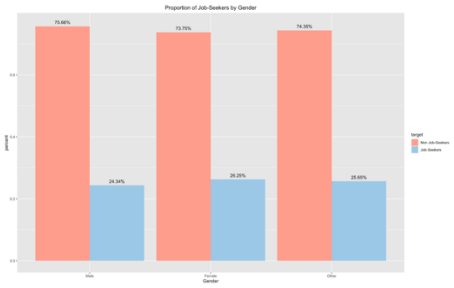
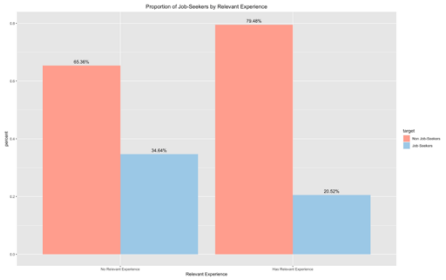
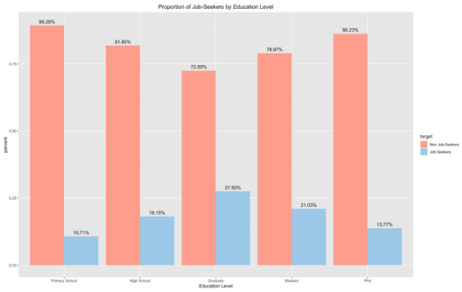
C1. More Exploratory Data Analysis

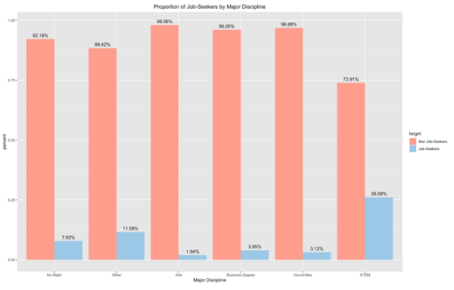
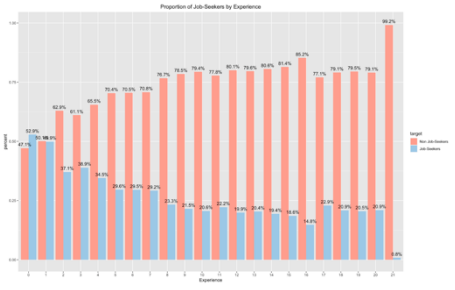
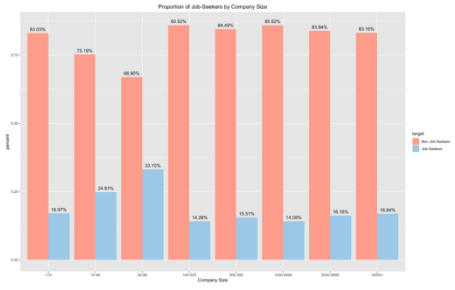
This section contains more insights from the exploratory data analysis for Job Seeker prediction.

Diagram	Insights
	<u>Training Hours vs Job Seeking</u> There is a higher proportion of job seekers with low training hours. And the number of employees willing to change jobs decreases as the number of training hours increases. Individuals with low training hours have limited skills and experience and will be less attractive to potential employers as employers prefer high skilled workers to increase productivity. Thus, individuals with low training hours face difficulty in changing jobs and securing employment and thus would have to seek jobs more actively. Moreover, individuals with low training hours may be less confident about their skills and more inclined to desperately and constantly apply for as many jobs as possible.

Click here to enter text.

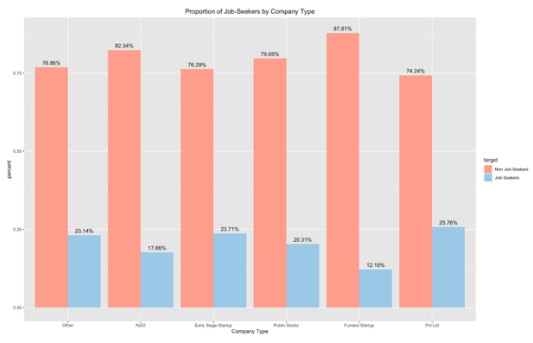
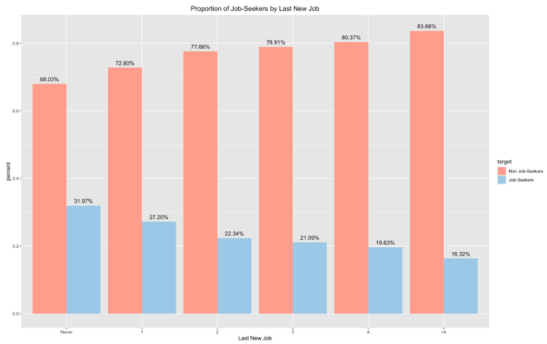
Univariate Analysis

Gender vs Job seeking	Relevant experience vs Job seeking	Education level vs Job seeking
		
The proportion of female job seekers is higher than male job-seekers due to gendered division of labour where women are still expected to take care of the caretaking responsibilities at home which causes women to take gaps between their employment. Workplace discrimination may also be a key factor preventing women from advancing in their careers. Thus, females may feel that they are being held back and thus may ought to actively seek in order to change to a new job that would allow them to advance in their careers.	Individuals with no relevant experience tend to actively seek jobs. Individuals with no relevant experience may find it difficult to change jobs in a competitive job market. Individuals with relevant experience would have an edge over individuals without relevant experience and thus recruiters would prefer those with the relevant experience. Due to lack of available positions, individuals with no relevant experience would have to constantly look for jobs.	A higher proportion of job seekers are made up of individuals who hold a graduate degree due to some industries like data science where a graduate degree is a minimum requirement. Thus, there would be greater competition among data scientist graduates to change jobs so they would have to actively seek jobs. A lower proportion of job seekers are made up of individuals with primary school education level as they would have limited job opportunities, lacking the skills and qualifications that may be a requirement in industries like data science. Since the government has support schemes in place for encouraging the current generation to pursue higher education, the dataset about individuals with primary school education level may be from the older generation who may be retired, who would not need and would not be able to actively seek employment.

Major discipline vs Job seeking	Experience vs Job seeking	Company size vs Job seeking
		

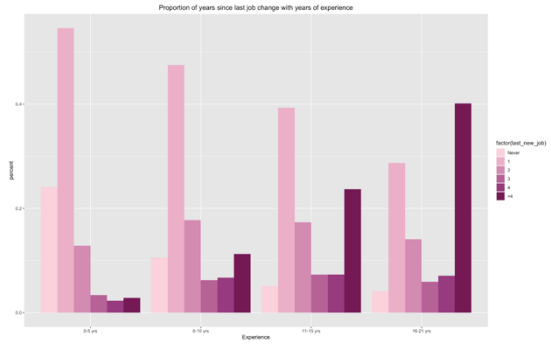
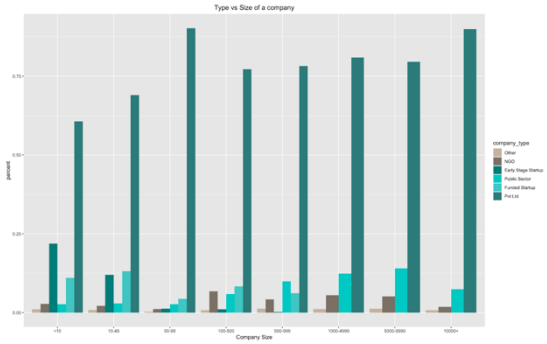
Click here to enter text.

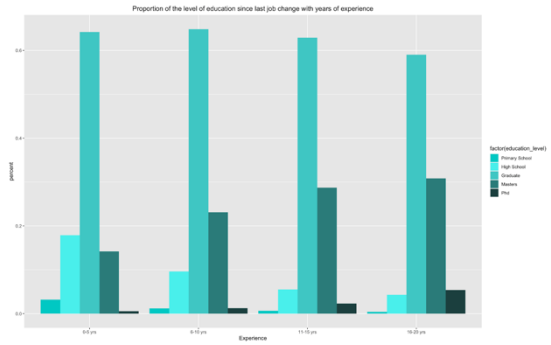
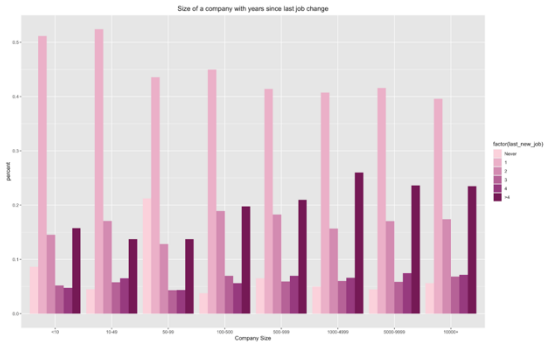
Higher proportion of individuals from STEM majors are actively seeking jobs compared to individuals from business, arts and humanities majors. Industries from STEM fields require specialised skills which are in high demand in the job market. In the technology industry, there is a greater preference for hiring STEM majors. As such there would be a higher proportion of individuals with STEM majors actively looking to explore career changes in data science. This is expected since the data is collected based on the data science industry.	Higher proportion of job seekers are made up of individuals with lower years of experience. Individuals with lesser years of experience would mostly be individuals of a younger age who will have a "hunger" for growth and career advancement. Thus, they will constantly look for new opportunities that offer higher salaries and more opportunities for success. Individuals with more experience will be seen as assets and will be highly valued by their current companies. Thus, they will be well taken care of with high employee benefits, convincing them to sustain in the company.	More individuals are willing to change jobs when the company size is low, in the range of 50-99. Small companies tend to have limited growth opportunities inhibiting individuals career development. Thus, there is a greater proportion of individuals actively seeking for greater growth potential.

Company type vs Job seeking	**Last new job vs Job seeking**
	
Most Individuals from Funded startup are not actively seeking employment as funded startups require a significant amount of time and resources. As such individuals involved in the funded startups will be highly invested in the success of the startup. Funded startups are more established than early startups with a growing customer base, requiring commitment from the individuals working for the funded startup. There is a higher proportion of job seekers from private limited companies. Private limited companies often have smaller company size and have limited resources which may make it difficult for employees to advance in their careers. Due to limited growth opportunities in private limited companies, employees may feel stagnant in their current role, thus would want to seek out a better job that would allow them to advance in their careers.	There is a high proportion of individuals who tend to job hop every 1 to 2 years. Job hoppers are expected to actively seek jobs. However, there is a larger proportion of job seekers that are made up of individuals who have never changed jobs before. This may be due to changes in the job market, encouraging such individuals to adapt to seek new job opportunities and find a job role that is high in demand. In search of a change, individuals who have never changed jobs before might also look for jobs out of boredom in the desire to explore new job opportunities and new work culture.

Click here to enter text.

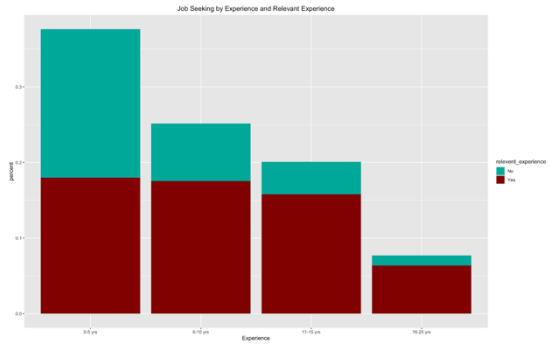
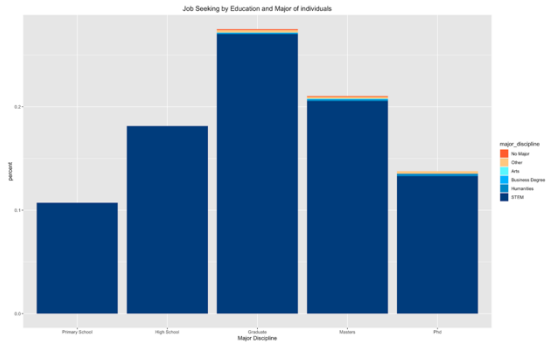
Bivariate Analysis

Experience vs Last new job	Company size vs Company Type
	
Individuals who have recently changed jobs within a year tend to have lesser years of experience which suggests that these individuals are most likely to be in the early stages of their career. Early stages of a career is a phase for most individuals to explore, experiment and find the best fit that matches their interest. Thus, individuals in the early stages of their career would be constantly looking for new job opportunities to upskill themselves.	It can be seen that there is no limit to the size of a private limited company. There are small private limited companies with less than 10 people and there are private limited companies with more than 10000 workers. However, in our dataset there is a higher proportion of private limited companies that have a company size in the range 50-99.

Education vs Experience	Last new job vs Company size
	
Individuals with lower levels of education have lesser experience. There is a greater proportion of individuals with high school education who have 0 to 5 years of experience. However, years of experience for individuals with graduate degrees has a wider range from 0 to 20 years.	It can be observed that most individuals from smaller company sizes tend to change jobs within a year gap. Most individuals from larger company sizes do not change jobs often. Larger companies typically have a high turnover rate which will incentivise individuals to stay in the company to enjoy the high profits of the company which may be translated through bonuses for the individuals.

Click here to enter text.

Multivariate Analysis

Job seeking by experience & relevant experience	Job seeking by major & education of individuals
	
Most of the job seekers have 0 to 15 years of relevant experience. Certain industries, like data science, tend to have a higher demand for entry level or mid-level positions, which will incentivise employees with less experience to change jobs. Additionally, economic conditions may influence the volatile job market which may influence employers to hire candidates with less experience in order to reduce costs. Thus, employees may use this as an opportunity to actively seek employment to change jobs with a hope to advance in their career.	Most of the individuals who are looking for jobs are STEM graduates. This is expected since the data is collected based on the data science industry. This is due to some industries like data science where a graduate degree is a minimum requirement. Thus, there would be greater competition among data scientist graduates who want to change jobs to advance in their career. Thus, data scientists would have to actively seek jobs.

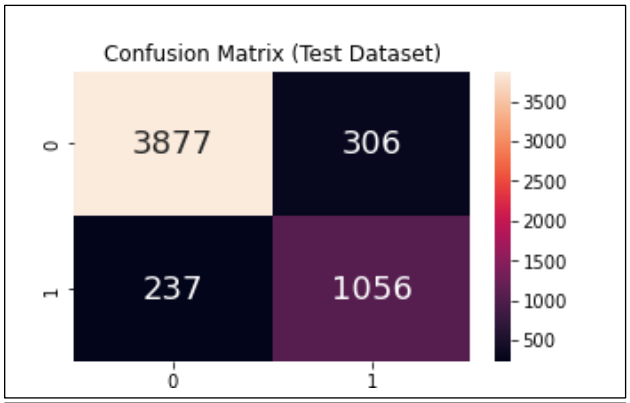
Click here to enter text.

C2. Detailed Model Performance Results

C2.1 Classification and Regression Tree [Full Dataset & Hyperparameter Tuning]

```
-----
-> Test Dataset
+-------------------------+------------+
| Metric                  |      Value |
+=========================+============+
| Classification Accuracy | 0.90084    |
+-------------------------+------------+
| True Positive Rate      | 0.816705   |
+-------------------------+------------+
| True Negative Rate      | 0.926847   |
+-------------------------+------------+
| False Positive Rate     | 0.0731532  |
+-------------------------+------------+
| False Negative Rate     | 0.183295   |
+-------------------------+------------+

Precision: 0.775330396475771
Recall: 0.816053364269141
f1_score: 0.7954802259887006
```




```
{'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': 10,
 'max_features': 20,
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'random_state': None,
 'splitter': 'best'}
```

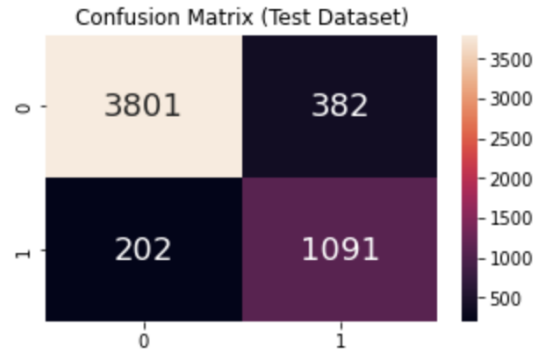

```
ROC-AUC-Score: 0.9286646924103917
```


Click here to enter text.

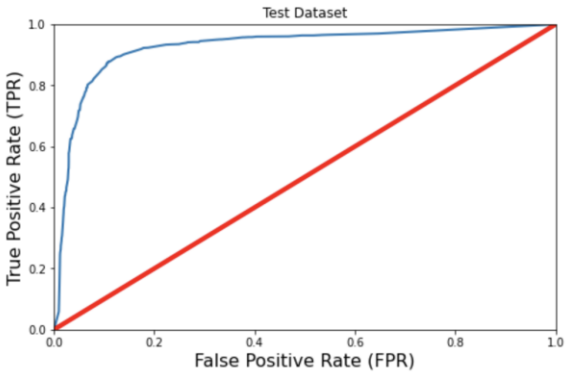
C2.2 Classification and Regression Tree [Selected Features & Hyperparameter Tuning]

Metric	Value
Classification Accuracy	0.893353
True Positive Rate	0.843774
True Negative Rate	0.908678
False Positive Rate	0.091322
False Negative Rate	0.156226

Precision: 0.7406653088934148
Recall: 0.8437741686001546
f1_score: 0.7888647866955893




```
{'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': 10,
 'max_features': 5,
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'random_state': None,
 'splitter': 'best'}
```



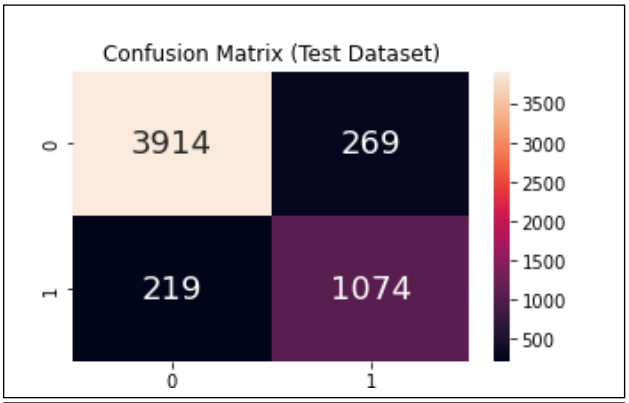
ROC-AUC-Score: 0.9266020586770857

Click here to enter text.

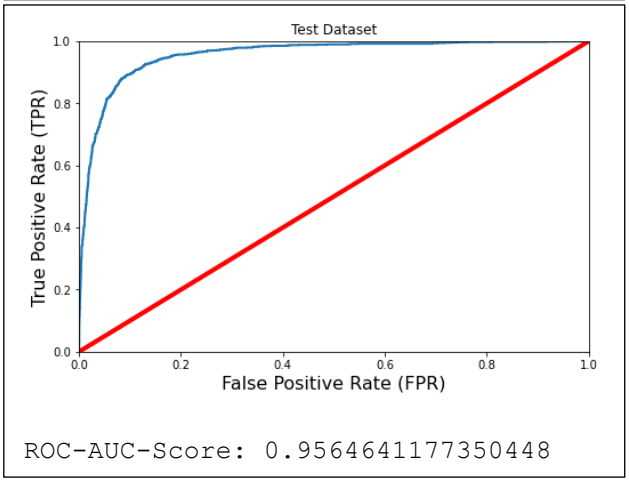
C2.3 Random Forest [Full Dataset & Hyperparameter Tuning]

```
-> Test Dataset
+-------------------------+-----------+
| Metric                  |     Value |
+=========================+===========+
| Classification Accuracy | 0.910884  |
+-------------------------+-----------+
| True Positive Rate      | 0.830626  |
+-------------------------+-----------+
| True Negative Rate      | 0.935692  |
+-------------------------+-----------+
| False Positive Rate     | 0.0643079 |
+-------------------------+-----------+
| False Negative Rate     | 0.169374  |
+-------------------------+-----------+

Precision: 0.7997021593447505
Recall: 0.8306264501160093
f1_score: 0.8148710166919576
```




```
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'sqrt',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 500,
 'n_jobs': None,
 'oob_score': True,
 'random_state': None,
 'verbose': 0,
 'warm_start': False}
```




```
ROC-AUC-Score: 0.9564641177350448
```

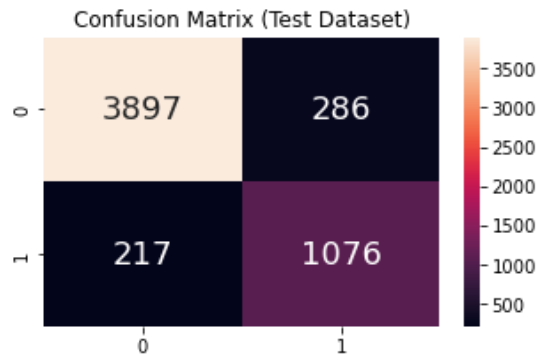

Click here to enter text.

C2.4 Random Forest [Selected Features & Hyperparameter Tuning]

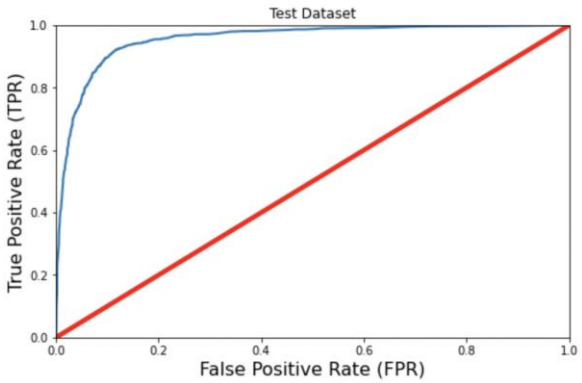
-> Test Dataset

Metric	Value
Classification Accuracy	0.908145
True Positive Rate	0.832173
True Negative Rate	0.931628
False Positive Rate	0.068372
False Negative Rate	0.167827

Precision: 0.7900146842878121
Recall: 0.8321732405259087
f1_score: 0.8105461393596988




```
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'entropy',
 'max_depth': None,
 'max_features': 'sqrt',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 500,
 'n_jobs': None,
 'oob_score': True,
 'random_state': None,
 'verbose': 0,
 'warm_start': False}
```



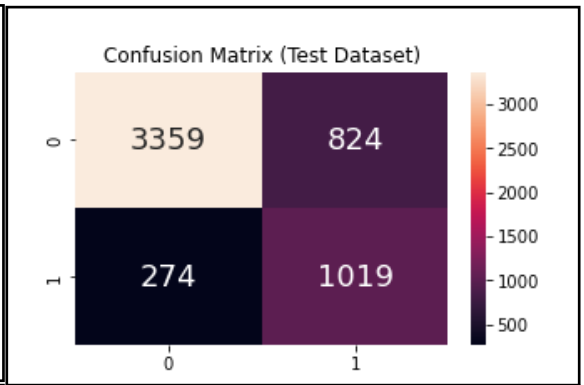
ROC-AUC-Score: 0.9554427849327157

Click here to enter text.

C2.5 Support Vector Classifier [Full Dataset & Hyperparameter Tuning]

```
-> Test Dataset
+-------------------------+-----------+
| Metric                  |     Value |
+=========================+===========+
| Classification Accuracy | 0.799489  |
+-------------------------+-----------+
| True Positive Rate      | 0.78809   |
+-------------------------+-----------+
| True Negative Rate      | 0.803012  |
+-------------------------+-----------+
| False Positive Rate     | 0.196988  |
+-------------------------+-----------+
| False Negative Rate     | 0.21191   |
+-------------------------+-----------+

Precision: 0.5529028757460662
Recall: 0.7880897138437741
f1_score: 0.6498724489795917
```




```
{'C': 1,
 'break_ties': False,
 'cache_size': 200,
 'class_weight': None,
 'coef0': 0.0,
 'decision_function_shape': 'ovr',
 'degree': 3,
 'gamma': 0.01,
 'kernel': 'rbf',
 'max_iter': -1,
 'probability': True,
 'random_state': None,
 'shrinking': True,
 'tol': 0.001,
 'verbose': False}
```

ROC-AUC-Score: 0.8850309108480372

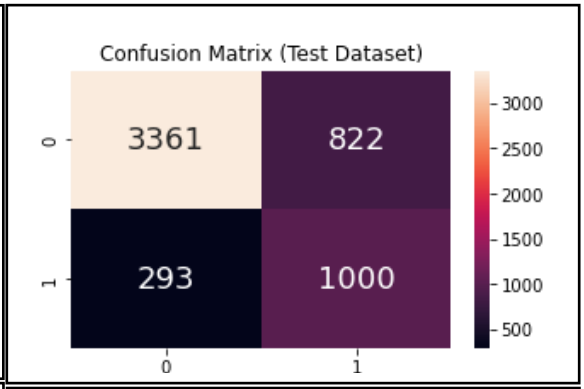
Click here to enter text.

C2.6 Support Vector Classifier [Selected Features & Hyperparameter Tuning]

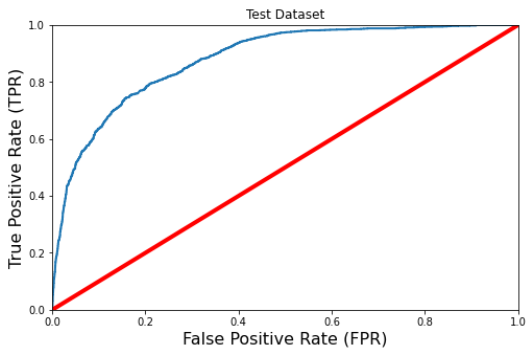
-> Test Dataset

Metric	Value
Classification Accuracy	0.796384
True Positive Rate	0.773395
True Negative Rate	0.80349
False Positive Rate	0.19651
False Negative Rate	0.226605

Precision: 0.5488474204171241
Recall: 0.7733952049497294
f1_score: 0.6420545746388443




```
{'C': 1,
 'break_ties': False,
 'cache_size': 200,
 'class_weight': None,
 'coef0': 0.0,
 'decision_function_shape': 'ovr',
 'degree': 3,
 'gamma': 0.01,
 'kernel': 'rbf',
 'max_iter': -1,
 'probability': True,
 'random_state': None,
 'shrinking': True,
 'tol': 0.001,
 'verbose': False}
```



ROC-AUC-Score: 0.8804616298541272

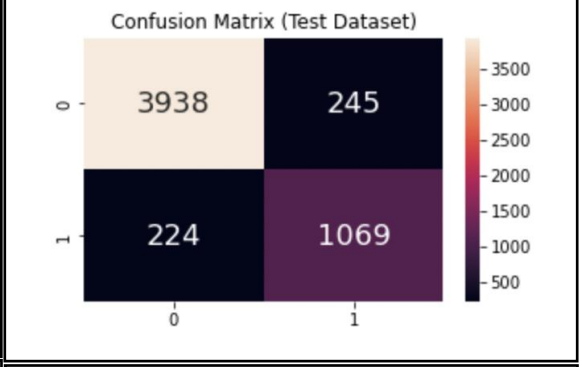
Click here to enter text.

C2.7 Extreme Gradient Boost [Full Dataset & Hyperparameter Tuning]

-> Test Dataset

Metric	Value
Classification Accuracy	0.914354
True Positive Rate	0.826759
True Negative Rate	0.94143
False Positive Rate	0.0585704
False Negative Rate	0.173241

Precision: 0.8135464231354642
Recall: 0.8267594740912606
f1_score: 0.8200997314921366




```
{'objective': 'binary:logistic',
 'use_label_encoder': False,
 'base_score': None,
 'booster': None,
 'callbacks': None,
 'colsample_bylevel': None,
 'colsample_bynode': None,
 'colsample_bytree': None,
 'early_stopping_rounds': None,
 'enable_categorical': False,
 'eval_metric': None,
 'feature_types': None,
 'gamma': None,
 'gpu_id': None,
 'grow_policy': None,
 'importance_type': None,
 'interaction_constraints': None,
 'learning_rate': None,
 'max_bin': None,
 'max_cat_threshold': None,
 'max_cat_to_onehot': None,
 'max_delta_step': None,
 'max_depth': None,
 'max_leaves': None,
 'min_child_weight': None,
 'missing': nan,
 'monotone_constraints': None,
 'n_estimators': 100,
 'n_jobs': None,
 'num_parallel_tree': None,
 'predictor': None,
 'random_state': None,
 'reg_alpha': None,
 'reg_lambda': None,
 'sampling_method': None,
 'scale_pos_weight': None,
 'subsample': None,
 'tree_method': None,
 'validate_parameters': None,
 'verbosity': None,
 'col_transformer__num_pipe__imputer__n_neighbors': 7,
 'estimator': XGBClassifier(base_score=None, booster=None, callbacks=None,
               colsample_bylevel=None, colsample_bynode=None,
               colsample_bytree=None, early_stopping_rounds=None,
               enable_categorical=False, eval_metric=None, feature_types=None,
               gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
               interaction_constraints=None, learning_rate=None, max_bin=None,
               max_cat_threshold=None, max_cat_to_onehot=None,
               max_delta_step=None, max_depth=None, max_leaves=None,
               min_child_weight=None, missing=nan, monotone_constraints=None,
               n_estimators=100, n_jobs=None, num_parallel_tree=None,
               predictor=None, random_state=None, ...),
 'estimator__colsample_bylevel': 0.6,
 'estimator__colsample_bytree': 0.6,
 'estimator__eta': 0.1,
 'estimator__gamma': 0,
 'estimator__max_depth': 28,
 'estimator__min_child_weight': 1,
 'estimator__n_estimators': 400,
 'estimator__subsample': 0.7}
```


ROC-AUC-Score: 0.9590470321536793

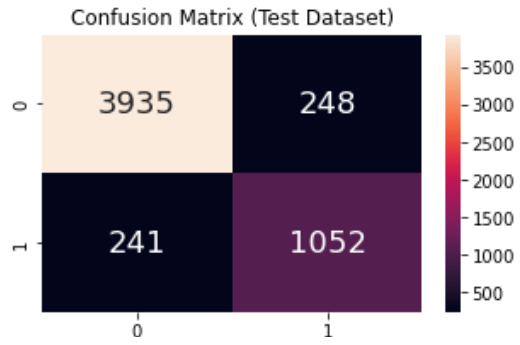
Click here to enter text.

C2.8 Extreme Gradient Boost [Selected Features & Hyperparameter Tuning]

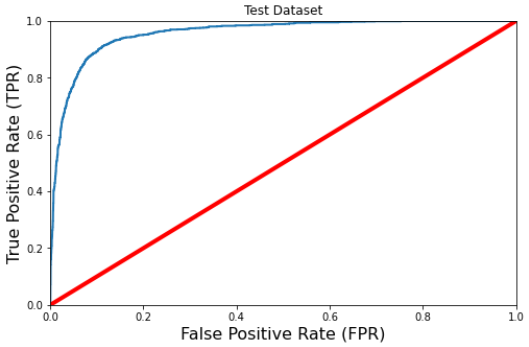
-> Test Dataset

Metric	Value
Classification Accuracy	0.910701
True Positive Rate	0.813612
True Negative Rate	0.940712
False Positive Rate	0.0592876
False Negative Rate	0.186388

Precision: 0.8092307692307692
Recall: 0.8136117556071152
f1_score: 0.8114153490165831



{'objective': 'binary:logistic',
'use_label_encoder': False,
'base_score': None,
'booster': None,
'callbacks': None,
'colsample_bylevel': None,
'colsample_bynode': None,
'colsample_bytree': None,
'early_stopping_rounds': None,
'enable_categorical': False,
'eval_metric': None,
'feature_types': None,
'gamma': None,
'gpu_id': None,
'grow_policy': None,
'importance_type': None,
'interaction_constraints': None,
'learning_rate': None,
'max_bin': None,
'max_cat_threshold': None,
'max_cat_to_onehot': None,
'max_delta_step': None,
'max_depth': None,
'max_leaves': None,
'min_child_weight': None,
'missing': nan,
'monotone_constraints': None,
'n_estimators': 100,
'n_jobs': None,
'num_parallel_tree': None,
'predictor': None,
'random_state': None,
'reg_alpha': None,
'reg_lambda': None,
'sampling_method': None,
'scale_pos_weight': None,
'subsample': None,
'tree_method': None,
'validate_parameters': None,
'verbosity': None,
'col_transformer__num_pipe__imputer__n_neighbors': 7,
'estimator': XGBClassifier(base_score=None, booster=None, callbacks=None,
colsample_bylevel=None, colsample_bynode=None,
colsample_bytree=None, early_stopping_rounds=None,
enable_categorical=False, eval_metric=None, feature_types=None,
gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
interaction_constraints=None, learning_rate=None, max_bin=None,
max_cat_threshold=None, max_cat_to_onehot=None,
max_delta_step=None, max_depth=None, max_leaves=None,
min_child_weight=None, missing=nan, monotone_constraints=None,
n_estimators=100, n_jobs=None, num_parallel_tree=None,
predictor=None, random_state=None, ...),
'estimator__eta': 0.1,
'estimator__gamma': 0,
'estimator__max_depth': 28,
'estimator__n_estimators': 400}



ROC-AUC-Score: 0.9568580260506426

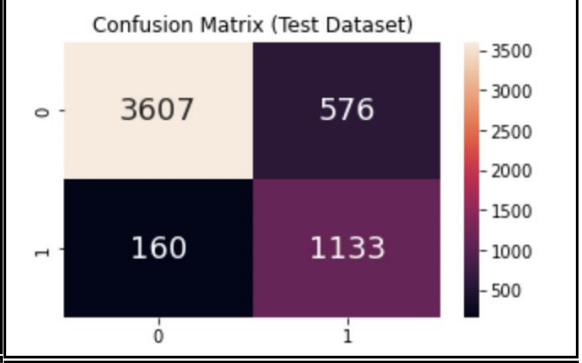
Click here to enter text.

C2.9 Logistic Regression [Full Dataset & Hyperparameter Tuning]

-> Test Dataset

Metric	Value
Classification Accuracy	0.865595
True Positive Rate	0.876257
True Negative Rate	0.8623
False Positive Rate	0.1377
False Negative Rate	0.123743

Precision: 0.66296079578701
Recall: 0.8762567672080434
f1_score: 0.7548301132578281




```
{'C': 1.0,
 'class_weight': None,
 'dual': False,
 'fit_intercept': True,
 'intercept_scaling': 1,
 'l1_ratio': None,
 'max_iter': 100,
 'multi_class': 'auto',
 'n_jobs': None,
 'penalty': 'l2',
 'random_state': None,
 'solver': 'lbfgs',
 'tol': 0.0001,
 'verbose': 0,
 'warm_start': False}
```

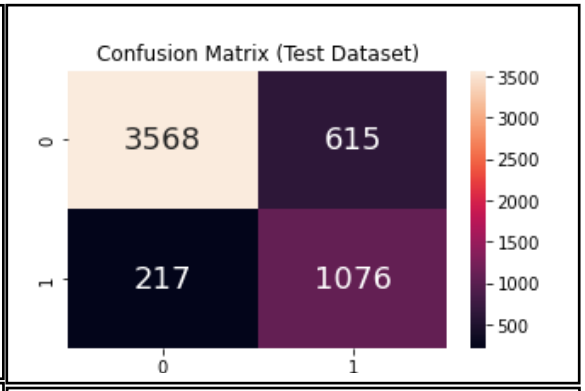
ROC-AUC-Score: 0.9273779498981163

Click here to enter text.

C2.10 Logistic Regression [Feature Selection & Hyperparameter Tuning]

```
-> Test Dataset
+-------------------------+-----------+
| Metric                  |     Value |
+=========================+===========+
| Classification Accuracy | 0.848064  |
+-------------------------+-----------+
| True Positive Rate      | 0.832173  |
+-------------------------+-----------+
| True Negative Rate      | 0.852976  |
+-------------------------+-----------+
| False Positive Rate     | 0.147024  |
+-------------------------+-----------+
| False Negative Rate     | 0.167827  |
+-------------------------+-----------+

Precision: 0.6363098758131284
Recall: 0.8321732405259087
f1_score: 0.7211796246648794
```




```
{'C': 1.0,
 'class_weight': None,
 'dual': False,
 'fit_intercept': True,
 'intercept_scaling': 1,
 'l1_ratio': None,
 'max_iter': 100,
 'multi_class': 'auto',
 'n_jobs': None,
 'penalty': 'l2',
 'random_state': None,
 'solver': 'lbfgs',
 'tol': 0.0001,
 'verbose': 0,
 'warm_start': False}
```


```
ROC-AUC-Score: 0.9165352560422539
```


Click here to enter text.