ST451 Bayesian Machine Learning

Assessed Coursework – Project

Candidate Number: 41466

Contents

Task 1 – Credit Risk Classification

Task 2 – Forecasting UK Exports

Task 1 – Credit Risk Classification

Introduction

Managing risk is vital for firms all over the world, especially banks. As banks lend out money in order to profit from interest, they cannot hold enough reserves to cover all of their deposit-holders, leaving them exposed to risk. As lending money is a main part of their business, it is very important for banks to know whether the people they are lending to are likely to pay back their debts on time. This makes the availability of a predictive model which can classify customers as 'good' or 'bad' risks a valuable tool.

Created by Professor Hans Hoffman, the 'German Credit Dataset' contains 607 observations of people who borrow credit from a bank [10]. Described by a set of 20 attributes, they are classified as a 'good' or a 'bad' credit risk to the bank. We aim to trial 8 different classification methods and compare their effectiveness. They are assessed through both their accuracy in predicting whether a person is a good or bad risk based on their attributes alone and their ability to avoid false positives (predicting a person is a good risk when they are bad). The candidate models we will test and evaluate are: Logistic regression (standard, ridge, stepwise & Bayesian), linear discriminant analysis, naïve Bayes, Gaussian process and support vector machine.

Attributes and Descriptive Analysis

The attributes we are given include the amount of money in the account, their credit history and some personal details with the full list available in Table 1 in the appendix. Many attributes are not useful for predicting good or bad credit risks; for example, the proportion of good or bad risks does not differ significantly between people who own a telephone and those who do not, or between foreign and German workers. However, as shown in Figure 1, each category of existing checking account has a different proportion of good and bad risks ranging from 64% of those with a negative checking account being good risks to 88% of those with a large amount in the checking account being a good risk.

Although not all attributes may hold strong predictive power, we cannot conclude they will not be useful altogether. The risk we could face when including all attributes is that our model becomes 'overfit'. This means that our model has been trained on information that is not useful in predicting credit risk and instead starts to attempt to describe the noise in our data rather than the underlying patterns we are interested in. However given the size of our dataset, most models are unlikely to overfit with just 20 attributes. We may want to reduce the number of predictors for the sake of simpler interpretation or only including significant associations in which case we can use a regularisation method such as 'Ridge' or 'Stepwise' model selection.

Methods and Models

First, we split our dataset at random into 670 observations that each candidate model is trained on. This means that the model sees all the attributes and whether each person is a good or bad risk in the training data. This allows the model to tune its parameters to understand which attributes correspond to people who are good or bad risks. Once the model

has trained its parameters, it is able to look at the attributes for a new person and make its best guess of the likelihood that they are a good risk, without seeing the true value. The remaining 330 observations in our dataset are called our test data. We will feed our models only the attributes of these 330 people and have the models predict whether they are good or bad risks. We can then compare what each model predicted to the true values thereby assessing how accurate each model will be at predicting good or bad risks when faced with new data.

Model Assessment

Unlike most datasets, the 'German Credit Dataset' offers instruction on how we should approach the classification problem and how we should assess our models. It provides us with a 'cost matrix' which describes how we should treat each possible prediction outcome. Although our target is credit risk, this cost is not monetary and is instead a guideline for how we should control our classifications. It says that any correct prediction (good or bad risk) receives no cost, a good risk which was predicted to be bad receives a cost of 1 and a bad risk which was predicted to be good receives a cost of 5 [9]. This essentially means that we penalise false positives (FP) five times as much as false negatives (FN) because offering credit to a risky customer is deemed more dangerous than refusing credit to safe customer.

This cost matrix also has implications for how we make our predictions. Given that we want to minimise the cost that our predictions incur, we are more interested in reducing the false positive rate than maintaining a high level of accuracy. Having written a function in Python to implement this assessment, we can input the predictions from any given model and it will return a score of the average cost incurred from each prediction.

If we assume *a priori* that people are equally likely to be good and bad credit risks, we can mathematically find the threshold probability that will minimise the expected cost in our predictions. This means that if we predict that every person with a predicted probability above this threshold is a good credit risk, we will minimise our cost in the long run. In equation (1) below, we use the formula for the optimal threshold under our assumptions to demonstrate why we should use a threshold probability of 0.833.

$$Optimal\ Threshold: \quad \theta^* = 1 - \frac{Cost_{FN}}{Cost_{FP} + Cost_{FN}} = 1 - \frac{1}{1+5} = \frac{5}{6} \approx 0.833 \qquad (1)$$

One could argue that the assumption of equal class priors is poor here given that most people are classed as good risks. However, we should try to avoid deciding our priors based on our training data as this is no longer *a priori* and practically, a small or moderate discrepancy in class priors is unlikely to make a big difference in minimising our cost. We could choose a threshold level once we have fit the models but in the aim of staying robust to new data, we would rather not choose our method of prediction based on the results and will instead maintain the same threshold across all models, keeping them more comparable.

Candidate Models

Logistic Regression

In a logistic regression, we model the chance that a person is a good risk by summing our chosen attributes, multiplying each by a value we chose in training and then transforming to

ensure we obtain a chance between 0 and 100 percent [4]. We choose the values that each attribute is multiplied by using a method called 'maximum likelihood estimation' [13]. This means that we choose the values that are most likely given the data that we have observed. Upon estimating the chance each test observation is a good risk, we predict that any over 0.833 are good risks and the rest are bad.

We can also use logistic regression with ridge regularisation meaning that during training, we attach a penalisation which shrinks our estimates and avoids overfitting [1], [14].

Another method of regularisation for logistic regression is called stepwise selection. This involves including attributes one by one only if their effect is deemed to be significant. This helps us avoid including factors that will overcomplicate our model without helping to classify risks.

Finally, Bayesian logistic regression approaches the same problem from a Bayesian perspective meaning we treat the values we are attempting to estimate as variable [11]. It also lets us incorporate prior beliefs on how different attributes affect credit risk and therefore often produces slightly different results to maximum likelihood estimation.

Linear Discriminant Analysis

Linear discriminant analysis works by looking at how spread out good and bad credit risks are across attributes and tries to draw a boundary between them while keeping the groups as far apart as possible. It also tries to minimise the variance in each group by keeping them as tight as possible and assumes that both good and bad credit risks follow the same type of distribution meaning they tend to spread out in a similar way [6].

Naïve Bayes

The naïve Bayes classifier tries to estimate how likely people are to be a good credit risk based on their attributes and gets its 'naïve' title from the assuming that all attributes (eg. job and housing) are independent. Although this is rarely true in most applications, the model is quite robust to all but strong violations of this assumption and is often an effective option [1].

Gaussian Process and Support Vector Machines

Gaussian process classifiers work by drawing smooth lines around our groups which are best fit to the 'kernel' we choose [2]. This 'kernel' is our chosen structure for how we want differences in attributes to mould our predictions. Support vector machines also use kernels but work by only looking at the people in opposite groups who have similar attributes and then drawing boundaries to maximise the distance between them [11].

Mean and Threshold Predictors

Just for comparison, we will show 'mean' and 'threshold' predictors which do not use the attributes at all and allow us to compare our models to a procedure that is hardly using any information from our training data. The mean predictor randomly predicts people to be good credit risks in the same proportion as the rest of our sample and the threshold predictor randomly predicts people to be good credit risks at the threshold rate that we determined earlier. If any of our models were to perform similarly or worse than these predictors, then we know that those methods are likely redundant for our task.

Results and Discussion

After training each model and assessing their predictions of the test data, we have a score for each model's accuracy (the proportion of its predictions that were correct) and its cost per prediction according to our cost matrix. While the accuracy can be a telling measure of a classifier's effectiveness, it is secondary to the cost per prediction in this scenario as our primary aim is to minimise the cost incurred by our predictions.

The best performing model was the full logistic regression using maximum likelihood estimation. Although the ridge and stepwise versions achieved a better fit to the training data (measured by a metric called AIC), this does not guarantee a better performance on out-of-sample predictions [8]. Listed in table 3, the logistic regression achieved an average prediction cost of 0.618 and also outperformed all other models on accuracy but it is worth noting that it was closely followed the linear discriminant analysis and Bayesian logistic regression methods which achieved average prediction costs of 0.621 and 0.633 respectively. The logistic regression's success compared with the other candidate models is likely due to the simple design of the model being well matched to the relatively low size and complexity of the data.

Notably, the Gaussian process and support vector machine classifiers achieved poor average prediction costs compared to alternatives at 0.694 and 0.709 respectively. Despite being considered powerful classification algorithms, their more complex decision boundaries were likely a poor match to the simple design of the data leading to a worse performance on data that the models were not trained on. Although these were worse than the other candidate models, all tested models clearly outperformed the mean and threshold predictors which achieved average prediction costs of 1.364 and 0.858 respectively.

Conclusion

In conclusion, the best performing model was the logistic regression closely followed by linear discriminant analysis and Bayesian logistic regression and it appears that the amount of credit, the amount of savings and the status of the existing checking account are among the most important attributes in determining a good or bad credit risk.

It is worth noting that many of these methods, especially the Gaussian process and support vector machines, often require more detailed tuning before they can be successful. Although the other methods could also be further enhanced for this task, it may be unfair to compare more complex models to methods that will work well almost 'out-of-the-box'. Therefore, further work could certainly be done to use each model closer to its full potential and to customise a different threshold for classification in each approach based on cross-validating to minimise prediction cost. Further research on estimating the true monetary cost of misclassification to then base our classification strategy on would also be interesting extension which would allow us to see the impact each model has in a more concrete way.

Crucially, it is important to note that our results do not show that logistic regressions are the inherent best way to classify credit risks but instead suggest that the most important consideration is picking a model that is well matched to the complexity of the data available.

Task 2 – Forecasting UK Exports

Introduction

International trade is an important part of any country's economy. In the UK, we heavily rely on buying goods and services from abroad and the Department of International Trade estimated in 2021 that 6.5 million UK jobs are supported by the exports we sell to other countries [3].

Time series analysis is the practice of examining how data changes over time and allows us to make principled and evidence-based forecasts on how things will look in the future. Exports are a constituent of gross domestic product (GDP) and the ability to produce reliable forecasts for how exports will change over time is a valuable tool for companies and central governments, for example, in determining consumer behaviour or in fiscal planning to support industry.

We have monthly data from January 1997 to February 2025 of UK goods exports to the rest of the world [12]. We are comparing two different approaches, the seasonal ARIMA model and a Gaussian process regression. A seasonal ARIMA model expresses exports as a combination of past values and allows us to incorporate seasonality and randomness into our predictions. Meanwhile, a Gaussian process regression is a flexible approach which allows us to fit a line of almost any shape through our data and offers a detailed output of the direction and uncertainty in our predictions. Our data is non-seasonally adjusted meaning we can make a more comprehensive judgement of how each approach can be used to handle trends, seasonality and randomness.

Methods and Models

As our data is time series, it no longer makes sense to randomly partition data into a training set and a test set. Instead, we will train our models on the first 326 months (January 1997 – February 2024) and test their accuracy in forecasting the past year (March 2024 – February 2025). We will then assess our models by measuring and comparing the root mean squared error and log prediction score in their forecasts. The root mean squared error is a measure of how close, on average, the forecasts were to the real values and the log prediction score combines this with assessing how accurate each model's uncertainty in their predictions were.

Seasonal ARIMA Model

A Seasonal Auto-Regressive Integrated Moving Average model (or S-ARIMA model) is a standard tool for modelling time series data and is applicable in many circumstances [7]. The 'auto-regressive' part is used to describe how export values depend on the previous few months. The 'moving average' part describes how 'noise', or randomness, affects the movement of export values and the 'integrated' part describes the trend in the data. Finally, the 'seasonal' component is used to describe how export values might follow a cyclic pattern. As our data is monthly, we could expect to see a seasonal pattern lasting 12 months and, as shown in Figure 6, export values show a strong correlation with the value exactly 12 months before. This confirms to us that we should use yearly seasonality for this model.

We can compare many different model forms and then pick the one that fits the training data the best (by using the AIC measure mentioned previously) [8]. The model with the strongest fit to the training data was a S-ARIMA(1,1,1)x(2,1,2)[12]. This means that we model our exports as following a straight trend over time, depending directly on the value and the noise from the month before and also depending on the value and noise from the same month for the two years prior.

During training, we will determine how much the previous months and years should be modelled as affecting each export value. We will use maximum likelihood estimation which means that we will pick the values that are most likely given the data that we have observed.

Gaussian Process Regression

A Gaussian process regression is a flexible way to fit a line through our data [2]. We will choose a 'kernel' that will describe how each point is correlated with the ones around it. A well-fitting kernel is very important for building a model and must be carefully considered given the data we have. We can use this to simultaneously model long or short-term trends, seasonal effects and noise. We can pick our model using a measure called the 'model evidence'. Unlike maximum likelihood estimation, this looks at the chance of observing our data for all possible values we could choose, allowing us to compare how different models fit our sample. Then for a chosen model, we optimise how previous exports should affect the exports that follow by choosing that values that produce the highest evidence.

Each of our models includes a structure to describe noise in the data. They also either attempt to describe a long or short-term trend. We can then extend these structures by choosing whether to add a part to describe the seasonal fluctuations in the export values. It is not always in our best interests to keep adding complexity to our structure as we risk overfitting. This is when our model starts describing the randomness in our dataset rather than the underlying trend, hurting inference and prediction.

Results and Discussion

Each model is trained on data until February 2024 and then produces a forecast for the next 12 months (March 2024 – February 2025). We then score their predictions against the true values by judging both the point predictions and the uncertainty in their forecasts through their root mean squared error and log prediction score.

The best performing model by root mean squared error (RMSE) was the short-term + seasonal + noise Gaussian process regression. It achieved a RMSE of £769 million, an error of only around 2.5% of the magnitude of the exports which are around £30 billion (units in the data are given in millions of pounds). However, this model achieved the poorest log prediction score out of any of the candidates with a score of 9.55 per prediction (a lower score is better). As shown in Figure 11, while the predictions are incredibly accurate, the prediction intervals are too wide, hurting its log prediction score.

Generally, the Gaussian process (GP) regressions seem to outperform the S-ARIMA model both in terms of RMSE and log prediction score showing that even though they are less common, they are clearly a powerful tool for forecasting time series data. As listed in Table 4, the best model in log predictive score was the long-term + seasonal + noise GP with a score

of 9.01 per prediction. It also had the highest model evidence of any of the GPs but only managed a RMSE of £1.25 billion. While this is still less than half that of the S-ARIMA model, it is notably higher than the models that use a short-term trend.

A good balance between RMSE and log prediction score is achieved by the short-term + noise model which achieved scores of £851 million and 9.09 respectively. Figure 9 shows that, although it fails to model any seasonality, it produces accurate forecasts and reasonable levels of uncertainty.

Conclusion

Overall, Gaussian process regressions generally outperformed traditional methods such as the S-ARIMA model in both RMSE and log prediction score. While different models performed best on each metric, the short-term + noise model achieved a good balance between both, showing that simpler models can often be better suited than complex ones, especially when dealing with datasets that are modest in their size and dimension.

However, making a comparison with the S-ARIMA model may not be fair in this situation. The S-ARIMA model assumes that volatility in export values is independent of time and this is not supported by the data. The plot of UK exports in Figure 5 shows that periods of high volatility seem to cluster together which is perhaps better described by an ARCH or GARCH model. Further work to develop our results could therefore be done on testing how these different models can better describe the volatility in exports and also on forecasting different periods throughout history, rather than just the last year.

Appendix

Figures 1-4: Proportion of Credit Risks by Attribute

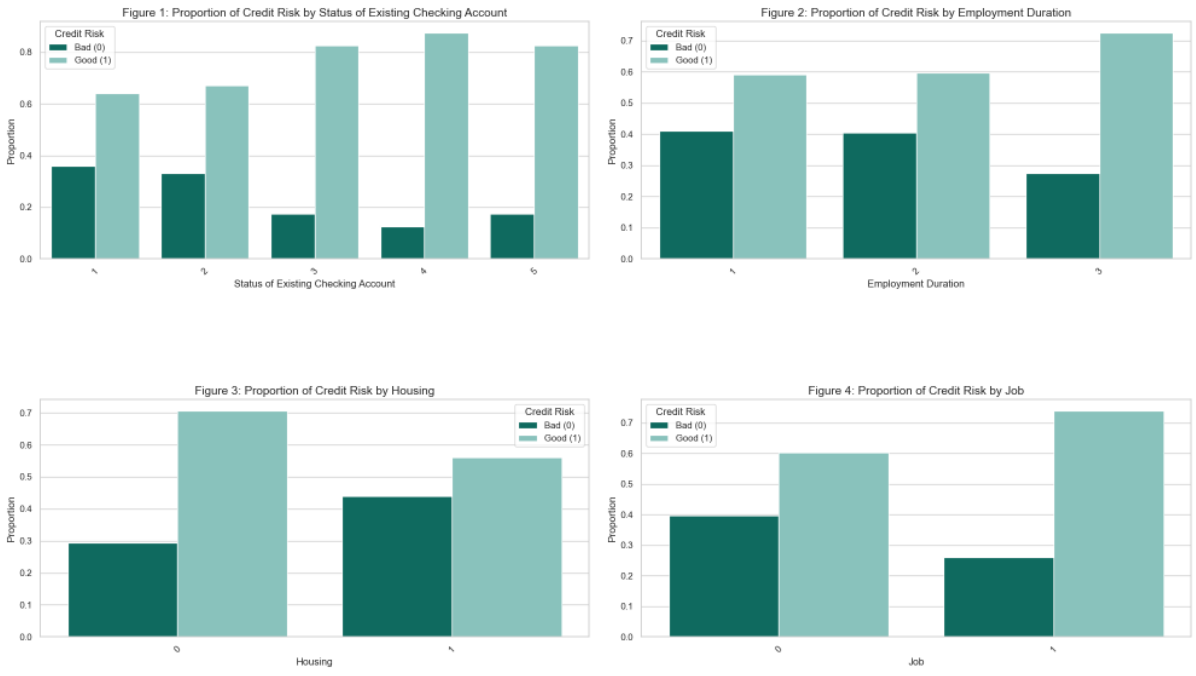


Table 1: Attributes and Short Description

Attribute	**Description**
Existing Checking	How much money is in the checking account (Categorical: Negative, small amount, large amount, no account.)
Duration	How many months since the account was opened
Credit History	Categorical: All paid duly, all at this bank paid duly, all paid eventually, some delays, critical credits existing
Purpose	List of uses for borrowing eg. car, education, business etc.
Credit Amount	Amount of money borrowed
Savings Account	Amount of money they have saved (categorical)
Employment	How long they have been employed (categorical)
Instalment Rate	Instalment rate in percentage of disposable income
Status and Sex	Sex and marital status (categorical)
Other Debtors	Categorical: none, co-applicant, guarantor
Present Residence	How long they have lived in their current residence
Property	Assets the person owns
Age	Age in years
Other Instalment Plans	Whether they have other instalment plans (Categorical: bank, stores, none)
Housing	Residential status (Categorical: rents, owns, for free)
Existing Credits	Number of existing credits at the bank
Job	Type of job (unemployed – highly qualified)
Number of Dependents	Number of people who depend on them financially
Own Telephone	Whether they have a telephone
Foreign Worker	Whether they are a foreign worker

Table 2: Cost Matrix

	Predicted Good Risk	**Predicted Bad Risk**
True Good Risk	Correct: 0 cost	False Negative: 1 cost
True Bad Risk	False Positive: 5 cost	Correct: 0 cost

Table 3: Classification Models and Average Error Scores

	Logistic Regression	Ridge Logistic	Stepwise Logistic	Bayesian Logistic	Linear Discriminant Analysis	Naïve Bayes	Gaussian Process	Support Vector Machine	Mean	Threshold
Average Error Score	0.618	0.655	0.633	0.633	0.621	0.645	0.694	0.709	1.364	0.858

Figure 5: UK World-wide Exports

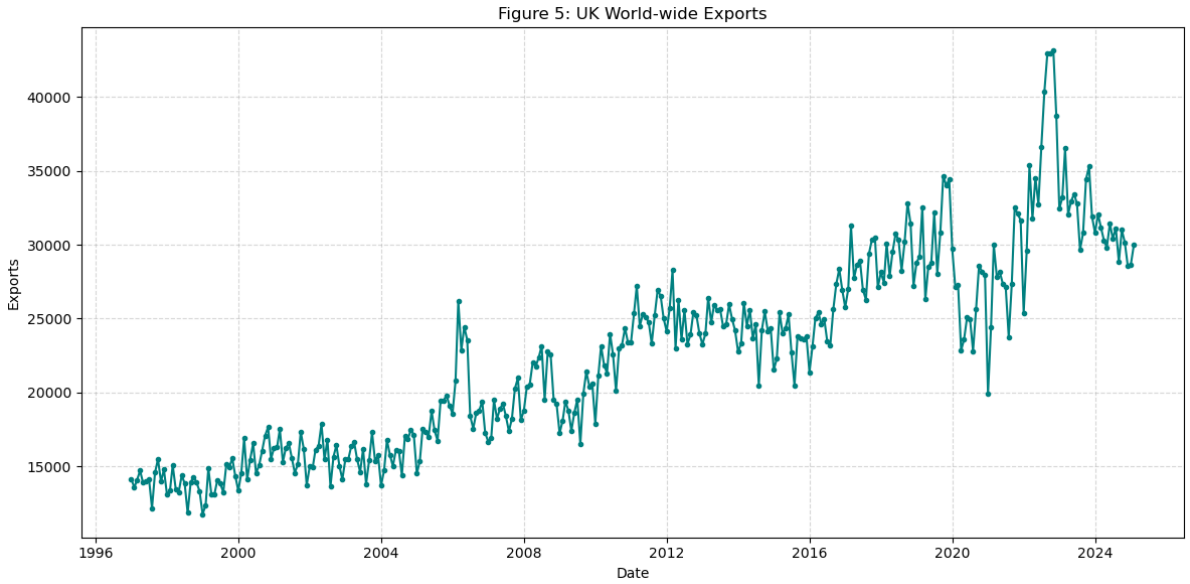
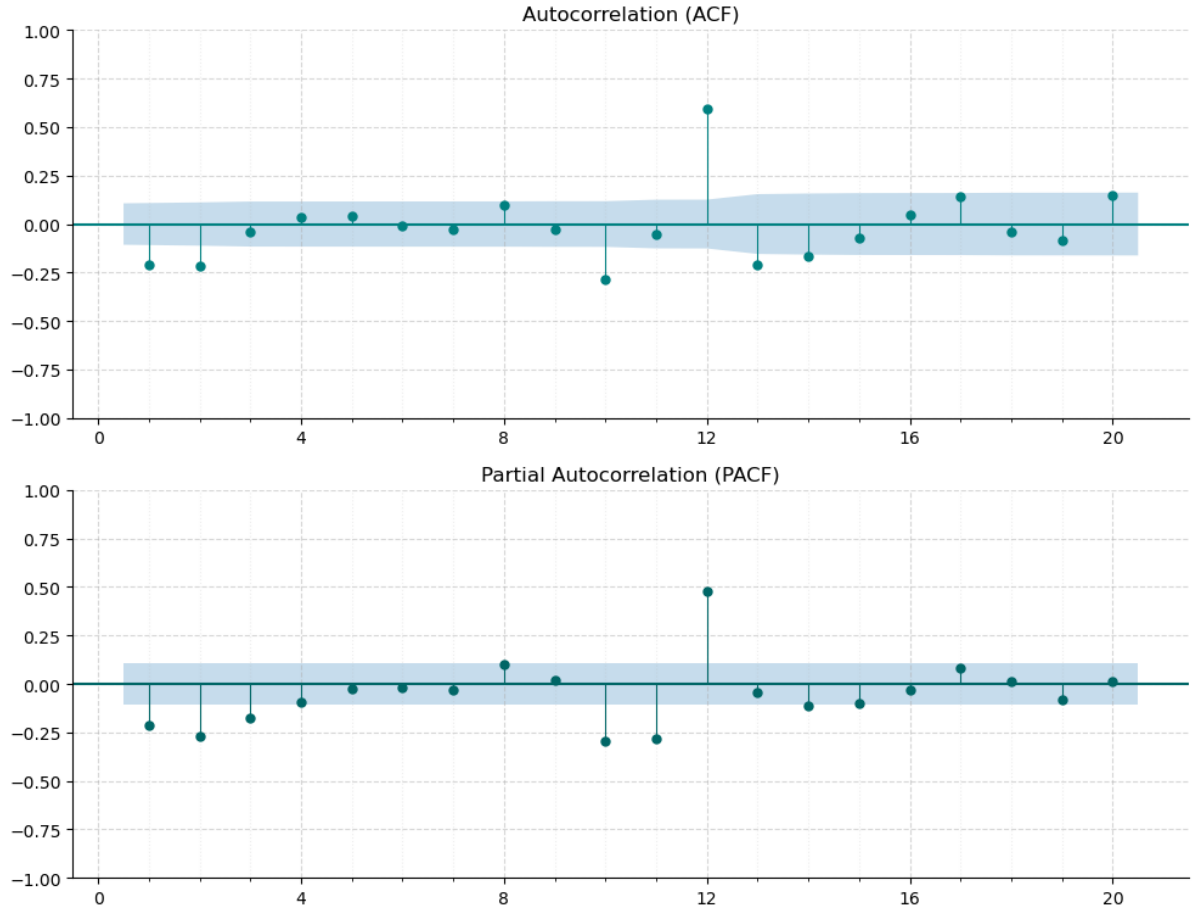
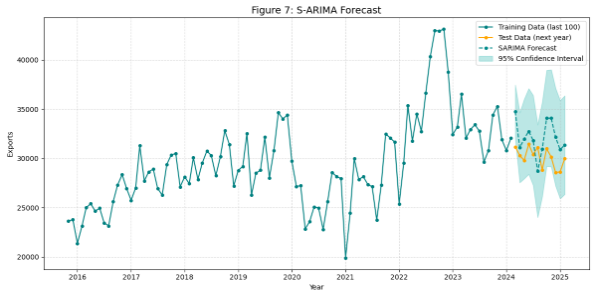
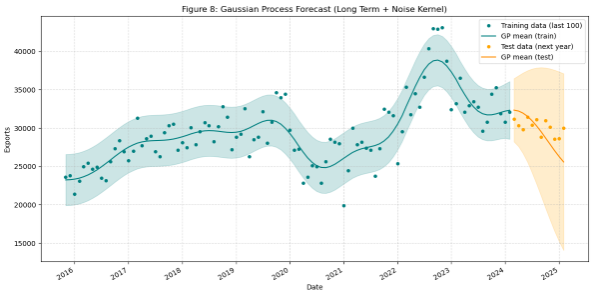


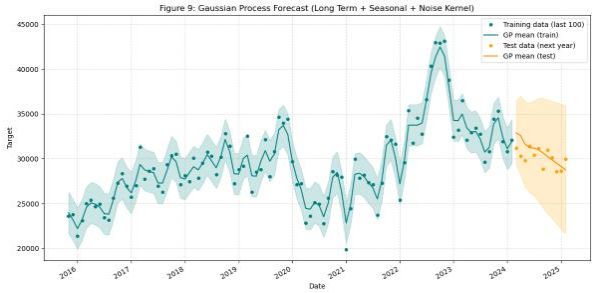
Figure 6: ACF & PACF for 1st Differenced Series

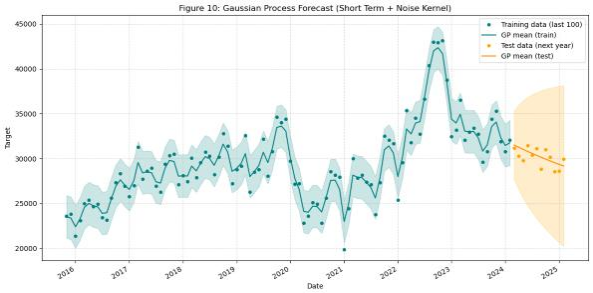


Figures 7-11: Forecasted Exports vs True Values (March 2024 – February 2025)









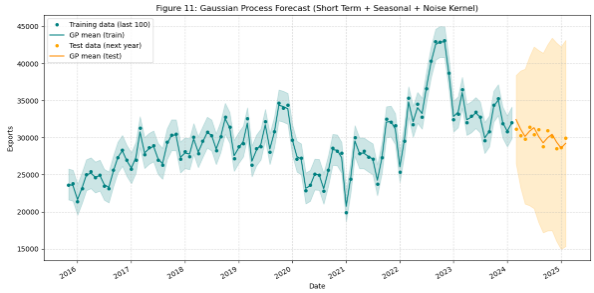


Table 4: Forecasting Models and Prediction Scores

	S-ARIMA	**Long Term + Noise GP**	**Long Term + Seasonal + Noise GP**	**Short Term + Noise GP**	**Short Term + Seasonal + Noise GP**
RMSE	2533	2007	1246	851.0	769.2
Log Prediction Score	9.363	9.286	9.009	9.088	9.554

References

[1] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[2] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press, 2006.

[3] Department for International Trade, *Evaluating the Impact of Exports on UK Jobs and Incomes*, London: Department for International Trade, Mar. 10, 2021. [Online]. Available: https://www.gov.uk/government/publications/evaluating-the-impact-of-exports-on-uk-jobs-and-incomes. [Accessed: May 8, 2025].

[4] D. R. Cox, “The regression analysis of binary sequences,” *J. Roy. Stat. Soc. Ser. B (Methodological)*, vol. 20, no. 2, pp. 215–232, Jul. 1958, doi: 10.1111/j.2517-6161.1958.tb00292.x.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Sec. 4.3, pp. 106–119, New York: Springer, 2008.

[6] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[7] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed., Hoboken, NJ: Wiley, 2015.

[8] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[9] H. Hoffmann, “Learning with Costly Features,” Ph.D. dissertation, Oregon State Univ., Corvallis, OR, 2009.

[10] H. Hofmann, “Statlog (German Credit Data),” *UCI Machine Learning Repository*, 1994. [Online]. Available: https://doi.org/10.24432/C5NC77.

[11] J. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 1999, pp. 61–74.

[12] Office for National Statistics, “Trade in Goods (T): WW: Exports: BOP: CP: NSA – LQAD,” *UK Trade Time Series (MRET)*, Feb. 13, 2025. [Online]. Available: https://www.ons.gov.uk/economy/nationalaccounts/balanceofpayments/timeseries/lqad/mret. [Accessed: May 8, 2025].

[13] R. A. Fisher, “On the interpretation of χ^2 from the standpoint of likelihood,” *J. Roy. Stat. Soc.*, vol. 85, no. 1, pp. 87–94, 1922.

[14] S. Seabold and J. Perktold, “Statsmodels: Econometric and Statistical Modeling with Python,” in *Proc. 9th Python in Science Conf. (SciPy 2010)*, Austin, TX, 2010, pp. 57–61.