Josh Besse

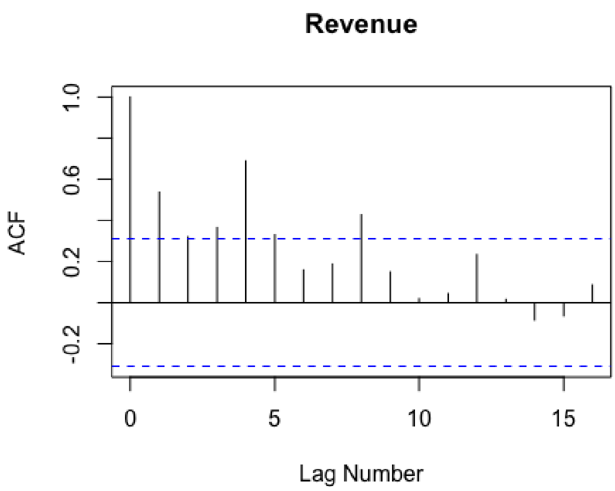
Netflix Forecasting

Introduction

Netflix is a streaming service that offers a wide variety of TV shows, movies, and documentaries. Netflix was founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California. It initially started as a DVD rental service; however, it has evolved into a subscription-based streaming service. It also started producing its own content in 2013, known as Netflix Originals. The data given is quarterly data from 2010 to 2019 and includes Netflix's total revenue, US GDP, US CPI, US unemployment, and US personal disposable income. The goal is to create a proficient model in forecasting Netflix's revenue, determine if Netflix producing its own content has been worthwhile in terms of eventual revenue, and verify if Netflix is recession-proof or not. To determine the "best" forecasting model, a multiple regression model, an exponential smoothing model, an ARIMA model, and a combination model will be employed. The performance of each model will be judged based on its mean squared error (MSE).

Results

When looking at the Autocorrelation Function (ACF), it suggests that the data has a trend and seasonal component. The declining nature of the ACF indicates the data has a trend, and the repeating spikes at every 4th lag indicates a quarterly seasonal pattern.



Thus, for the multiple regression, three dummy variables (Q1, Q2, and Q3) were created to see if the seasonality is really significant. In this case, quarter 4 is left out, which means it serves as the average. Also, a trend variable was created to account for the long-term trend. Three multiple regression models were created: all independent variables (GDP, CPI, unemployment, personal disposable income, Q1, Q2, Q3, and trend), all significant variables from the first model (GDP, Q1, Q2, Q3), and a stepwise regression. The backward stepwise regression and mixed stepwise regression (GDP, unemployment, Q1, Q2, Q3) proved to be the "best" model with a MSE of 13023646.

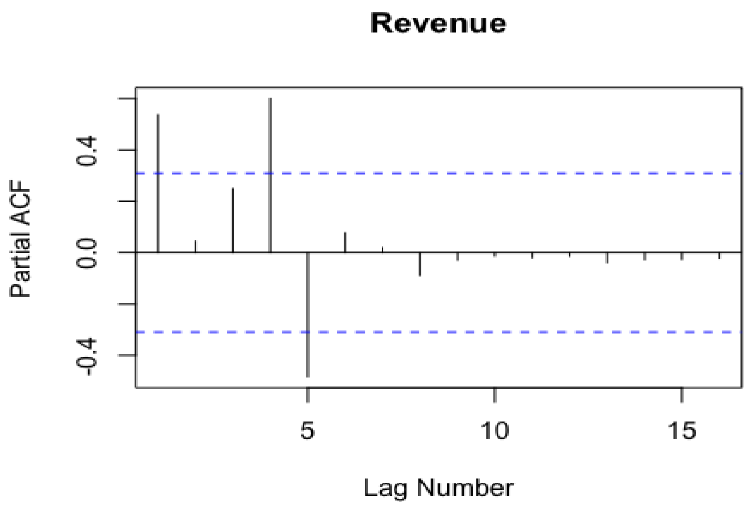
```
Residuals:
    Min      1Q  Median      3Q     Max 
-3200.3  -452.7     4.8   596.5  2664.4 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -7.477e+04  1.378e+04  -5.425 7.02e-06 ***
GDP_Billions         3.803e+00  5.884e-01   6.463 3.85e-07 ***
US_Unemployment_Rate  2.206e+03  5.317e+02   4.148 0.000254 ***
Q1                  -4.278e+03  5.716e+02  -7.485 2.42e-08 ***
Q2                  -3.033e+03  5.669e+02  -5.349 8.69e-06 ***
Q3                  -1.597e+03  5.654e+02  -2.825 0.008336 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1198 on 30 degrees of freedom
Multiple R-squared:  0.9003,	Adjusted R-squared:  0.8837 
F-statistic: 54.17 on 5 and 30 DF,  p-value: 4.073e-14
```

The ANOVA table p-value (<.001) indicates that the entire model is significant, and all parameters are significant. The assumptions for OLS are met. The error terms are normally distributed. There are no error terms greater than three standard deviations away from the mean, and the number of error terms that are positive is approximately equal to the number of error terms that are negative. The error terms have a constant variance, meaning they are homoscedastic. There does not appear to be a pattern to the error terms. The error terms are statistically independent. The Durbin-Watson statistic is 1.7469, which is within the range of a “good” DW statistic and means there is no problem with autocorrelation. The strengths of the multiple regression model are that it can incorporate various independent variables, and it is easy to interpret the relationship between variables. However, it is not as good at following the data compared to some of the stochastic models.

Since there is a quarterly seasonal pattern to the data, three exponential smoothing models were created: simple seasonal model, Winters’ additive model, and Winters’ multiplicative model. Exponential smoothing models follow trend and seasonality of data well. The Winters’ multiplicative model turned out to be the most effective exponential smoothing model with a MSE of 244378.1.

```
Forecast method: Holt-Winters' multiplicative method

Model Information:
Holt-Winters' multiplicative method 

Call:
 hw(y = train_series, seasonal = "multiplicative") 

  Smoothing parameters:
    alpha = 0.8704 
    beta  = 0.2702 
    gamma = 1e-04 

  Initial states:
    l = 995.8735 
    b = 228.5352 
    s = 1.5498 1.1958 0.8264 0.428

  sigma:  0.0439

     AIC     AICc      BIC 
490.3645 497.2876 504.6162 

Error measures:
                  ME     RMSE      MAE         MPE     MAPE       MASE      ACF1
Training set 25.3647 103.3606 79.18478 -0.06212598 2.718116 0.07569716 0.5993171
```

Considering that the data has a long-term trend and is seasonal, the ACF shows that the data is autocorrelated, and there are spikes in the PACF, an ARIMA model requires differencing, seasonal components, and AR and MA components. ARIMA models can model a wide range of data, but it is difficult to understand what exactly the model is doing ("black box").



Using the auto.arima function from the forecast package, the best fit ARIMA model is an ARIMA(0,1,1)(0,1,0). The ARIMA model has a MSE of 2549634.4.

```
Series: train_series
ARIMA(0,1,1)(0,1,0)[4]

Coefficients:
          ma1
      -0.4645
s.e.   0.1846

sigma^2 = 425195:  log likelihood = -244.49
AIC=492.97   AICc=493.4   BIC=495.84

Training set error measures:
                   ME     RMSE      MAE       MPE    MAPE      MASE        ACF1
Training set 162.7464 595.2557 437.9973 -2.918654 12.8419 0.4187061 0.005509617
```

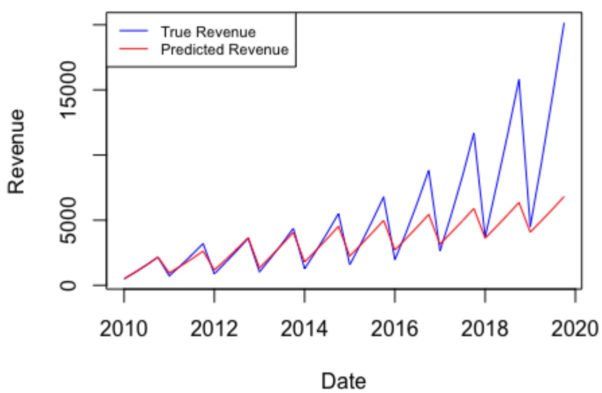
The two combination models utilized were multiple regression and exponential smoothing and multiple regression and ARIMA. Combination models combine the strengths of causal (regression) and stochastic (exponential smoothing and ARIMA) models but are more complex to interpret and implement. When running a regression with revenue as the dependent variable and the predicted values from the multiple regression and exponential smoothing models as the independent variables, the constant is not significant. This means the two models can be combined. When running the same regression without a constant, the results indicate that the contribution of the multiple regression model to the prediction of revenue is negligible. When performing the same process with the ARIMA model instead of the exponential smoothing model, the contribution of the multiple regression model to the prediction of revenue is again negligible.

“Best” Model

The “best” forecasting model is the Winters’ Multiplicative model. Out of all the models, it has the lowest MSE of 244378.1. The predictions it gave for 2019 Q1, 2019 Q2, 2019 Q3, and 2019 Q4 are 4609.245, 9361.874, 14217.100, and 19296.064. The forecasts indicate that Netflix’s revenue will continue to increase. They also indicate that quarter 1 will continue to be the worst quarter of the year for revenue and quarter 4 will continue to be the best quarter of the year for revenue.

Own Content Analysis

To determine if producing its own content is worthwhile for Netflix, intervention analysis can be employed. In order to perform intervention analysis, a variable that consists of revenue data up until the end of 2012 was created. Then, an ARIMA model was performed on that variable to predict what revenue would have looked like if Netflix didn’t start producing its own content.



There is a significant difference in the predicted revenue and true revenue. This indicates that Netflix producing its own content has been worthwhile, and it should continue production in the future.

Recession-Proof Analysis

To verify the claim that Netflix is recession-proof, an ARIMA model with revenue as the dependent variable and GDP, CPI, unemployment, and personal disposable income as the independent variables was utilized.

```
Series: data_series
Regression with ARIMA(0,0,0)(0,1,0)[4] errors

Coefficients:
       xreg1     xreg2      xreg3   xreg4
      3.0654   64.5698  1610.4179  0.1274
s.e.  0.5146  113.3578   355.7077  0.5425

sigma^2 = 633685:  log likelihood = -289.42
AIC=588.83   AICc=590.83   BIC=596.75

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 0.4460695 712.003 504.4601 -9.417004 16.07341 0.4167745 -0.01609048
```

The coefficient for GDP is 3.0654 and is significant, which suggests that an increase in GDP is associated with an increase in Netflix’s revenue. This might mean that Netflix’s revenue is not entirely immune to economic fluctuations. The coefficient for CPI is 64.5698 and is not statistically significant. This implies that changes in the CPI don't have a statistically significant association with Netflix's revenue and that Netflix's revenue is not sensitive to price changes in the short term. The coefficient for the unemployment rate is 1610.4179 and is statistically significant, indicating that there is relationship between unemployment rates and Netflix's revenue. The coefficient for personal disposable income is 0.1274 and is not significant. This suggests that changes in personal disposable income don't have a statistically significant association with Netflix's revenue. Overall, the model does not provide a definitive answer to whether Netflix is recession-proof or not, but it does suggest that Netflix's revenue is not highly sensitive to certain economic downturn indicators like CPI.