# Statistics for Data Analytics: Report on Time Series Analysis for Irish Car Registrations and Logistics Regression for Credit Institution Loan Default on Customers


Viplav Vijay Gadewar
*Statistics for Data Analytics*
*Master of Science in Data Analytics*
*National College of Ireland*
*Dublin, Ireland*
*x21164274@student.ncirl.ie*



***Abstract*—People buying cars and people utilizing credit cards for day-to-day expenditures are increasing at a rapid rate as time passes. This work focuses on two statistical models: time series analysis of automobile registration for a certain time period and use of Logistic Regression to predict credit card debt default.**



***Index Terms*—Time Series Analysis, Logistic Regression, Assumptions, ARIMA, Descriptive Statistics**


## I. Time Series Analysis

### A. *Introduction*

We investigate data that is not cross-sectional but in the time style format, i.e. data that was gathered on a regular basis over a period of time, using a statistical approach known as time series analysis. Annual, half-yearly, quarterly, monthly, weekly, daily, hourly, or at exact regular intervals, time-series data can be collected. We try to forecast, comprehend the trend, and check for seasonality by studying such time-series data.

### B. *Problem Description*

People are buying automobiles in increasing numbers all across the world, and this trend is positive, growing day by day as people's earning power grows. As a result, the number of cars registered in Ireland has grown. This study will use a time series analysis approach to forecast six periods in the future based on current automotive registration data.

### C. *Description of Dataset*

The time series data collected from the Central Statistics Office of Ireland consists of two columns: the date of registration in YYYY MM format from January 1995 to January 2022, and the count, which indicates the number of cars registered in that month. There are 325 observations in all. The structure of the data frame is shown in fig 1.

```
> car <- read.csv("C://Users//Viplav//Downloads//CarRegistrations.csv", header = FALSE, col.names = c
+                 ('time','count'), fileEncoding="UTF-8-BOM")
> #conversion to time series data
> carreg <- ts(car, start=c(1995,1), frequency = 12)
> dim(carreg)
[1] 325   2
```

Fig. 1. Data Description

We may choose the optimal prediction method by considering the time-series trend. A time-series can have none (stationary data), all, any one, or a combination of the following irregular patterns, notably trends, cyclic patterns, and seasonality variations. Time-series data items are what they're called.

- *1. Trend* -
  The trend is that each split within a dataset is a continuous timeline with no specified interval. Null, negative, or positive trends are all possibilities.
- *2. Seasonality* -
  Data is collected at predetermined times. The patterns are predictable and repeat themselves, swinging up and down.
- *3. Cyclical* -
  There are no defined intervals in it. The duration is not set and might be short or long.
- *4. Irregularity* -
  This can be used to depict unexpected brief events.

### D. *Seasonal Plot and Seasonal Subsidiaries Plot*

Raw time series data is shown in the fig. 2 below.

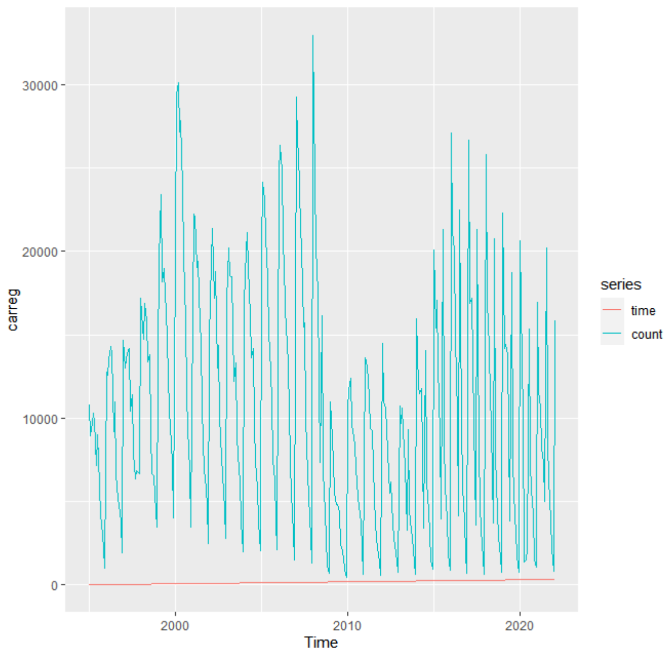


Fig. 2. Plot for Raw Data

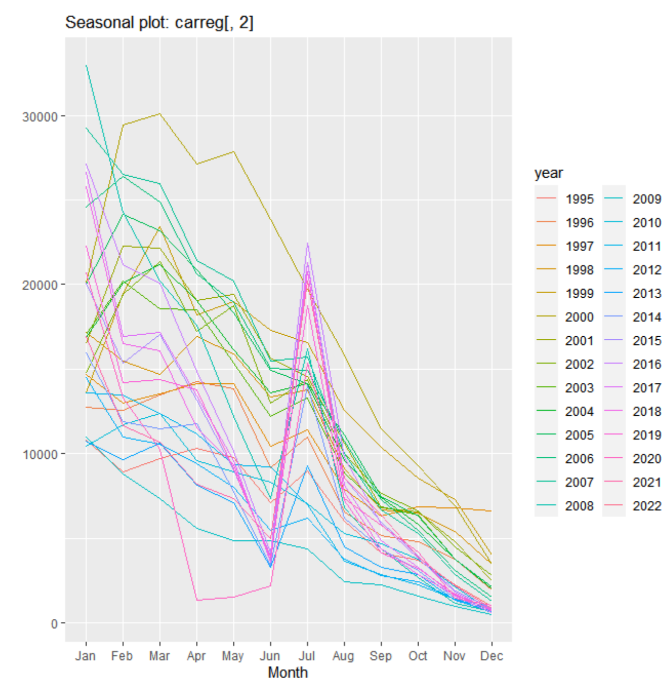


Fig. 3. Plot for Seasonal Data

The trend is at its height in the month of June, hence the data is seasonal, as seen in fig 3.

The data for each decade is put together throughout the months that follow. By aggregating data for a month or a specified period for all years and putting a horizontal line across them that represents the mean value for that month, the ggsubseriesplot() function in R creates mini graphs.

### E. Model building

#### 1) Simple time series model:

- Mean Model:
  By averaging certain prior data, the mean model forecasted the values for future months. The root mean square error (RMSE) of 7164.516 is shown in Figure 4.

```
Error measures:
                        ME     RMSE      MAE       MPE    MAPE     MASE      ACF1
Training set -5.254505e-13 7164.516 5969.178 -134.3288 164.893 2.732633 0.5691887

Forecasts:
         Point Forecast    Lo 80    Hi 80     Lo 95    Hi 95
Feb 2022       10510.54 1281.733 19739.35 -3627.737 24648.82
Mar 2022       10510.54 1281.733 19739.35 -3627.737 24648.82
Apr 2022       10510.54 1281.733 19739.35 -3627.737 24648.82
May 2022       10510.54 1281.733 19739.35 -3627.737 24648.82
```

Fig. 4. Mean Model Result

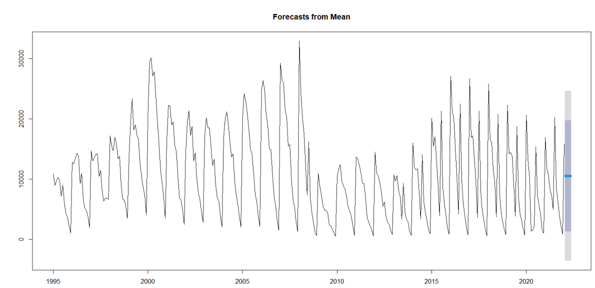


Fig. 5. Mean Model graph

- Seasonal Naïve Method:
  This method predicts the same-season value from the prior year. The seasonal naïve model is one of the easiest to use and delivers the most accurate results. Figure 6 shows the findings for our data, with an RMSE of 3475.135.

```
Error measures:
                   ME     RMSE      MAE       MPE     MAPE MASE      ACF1
Training set 57.47284 3475.135 2184.406 -9.056944 29.09186    1 0.7196226

Forecasts:
         Point Forecast    Lo 80    Hi 80   Lo 95    Hi 95
Feb 2022          11672 7218.435 16125.56 4860.86 18483.14
Mar 2022          10672 6218.435 15125.56 3860.86 17483.14
Apr 2022           8214 3760.435 12667.56 1402.86 15025.14
```

Fig. 6. Seasonal naive accuracy result

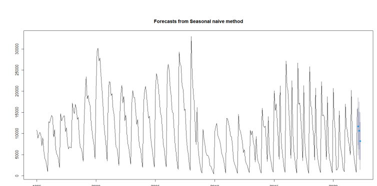

Fig. 7. Seasonal naive model graph

*2) **Exponential Smoothing**:* Exponential smoothing is a time series forecasting technique for univariate data. The model explicitly uses an exponentially decreasing weight for prior observations, similar to how exponential smoothing forecasting systems use a weighted sum of previous data to produce predictions.

- Holt-Winters Model:
  The Holt-Winters methodology is a prominent forecasting method that takes both trend and seasonality into consideration. Figure 8 shows the RMSE for holt-winter, which is 2604.6.

```
Error measures:
                    ME   RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -226.5401 2604.6 1820.253 -2.517114 32.13705 0.8332943 0.2135645

Forecasts:
         Point Forecast    Lo 80    Hi 80     Lo 95    Hi 95
Feb 2022      13335.089 9911.832 16758.35 8099.6697 18570.51
Mar 2022      12457.654 8634.552 16280.76 6610.7241 18304.58
Apr 2022       8724.431 4535.599 12913.26 2318.1644 15130.70
May 2022       7374.283 2845.540 11903.03  448.1682 14300.40
```

Fig. 8. Holt winter accuracy result

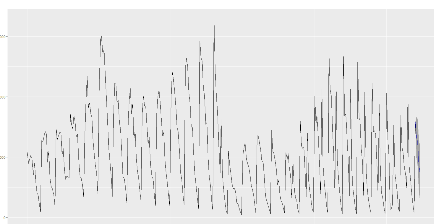

Fig. 9. Holt winter model graph

- ZZZ model ETS:
  The default value of ZZZ guarantees that all items are picked based on the informative criterion, which dynamically determines the error type, Pattern type, and Season type. The RMSE for this model is 2549.813 and the details are shown in Figure 10.

```
     AIC     AICc      BIC
6752.396 6754.389 6816.721

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE     ACF1
Training set -354.3682 2549.813 1540.122 -10.35789 20.12937 0.7050532 0.293539
```

Fig. 10. ZZZ accuracy result

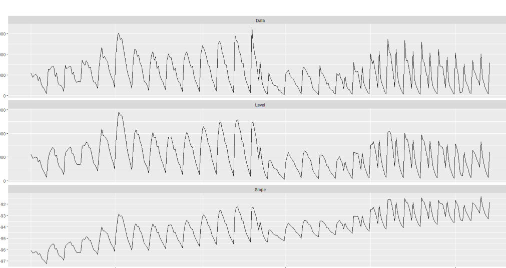

Fig. 11. ZZZ model graph

*3) **ARIMA Model**:* ARIMA is an acronym for Auto Regressive Integrated Moving Average. ARIMA models are statistical models used to evaluate and predict time series data. They can capture a range of common temporal patterns in time series data. It is tailored to a set of specified time-series patterns, and as a result, it provides a simple yet effective method for forecasting correct time series. It's a more advanced version of the Auto Regressive Moving Average with integration. In addition, we produced ACF and PACF plots as shown in figs 12 and 13 below.

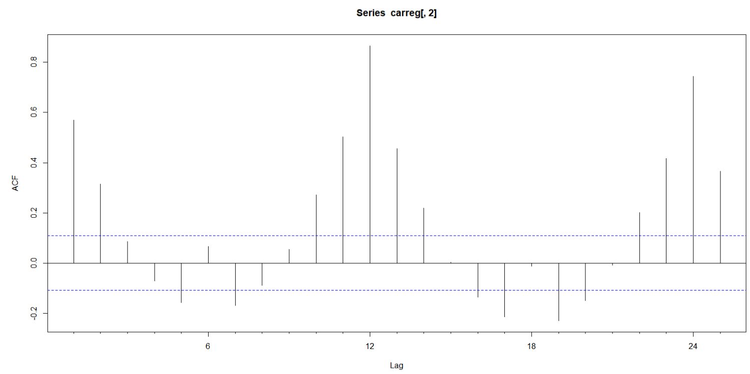


Fig. 12. ACF Model

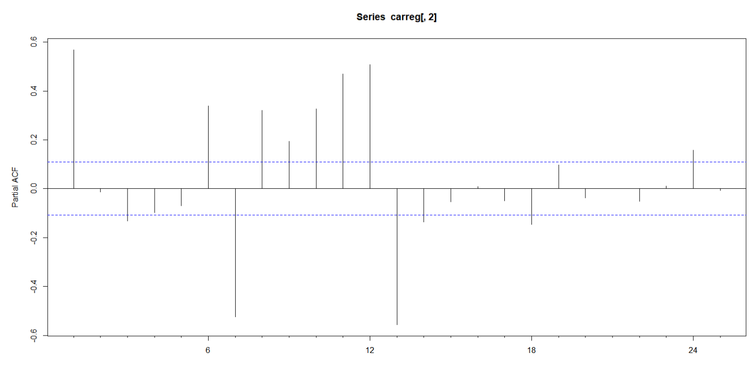


Fig. 13. PACF Model

The ARIMA model's accuracy/RMSE is determined to be 2247.616 as seen in the figure below.

```
> summary(arima.fit1)
Series: carreg[, 2]
ARIMA(1,0,1)(1,1,2)[12]

Coefficients:
         ar1      ma1    sar1     sma1     sma2
      0.8205  -0.2133  0.6527  -0.8451  -0.0334
s.e.  0.0457   0.0809  0.1292   0.1406   0.0796

sigma^2 = 5330611:  log likelihood = -2868
AIC=5748.01   AICc=5748.28   BIC=5770.49

Training set error measures:
                   ME     RMSE     MAE       MPE    MAPE      MASE       ACF1
Training set 21.46117 2247.616 1315.65 -2.557833 18.5022 0.6022919 0.01119676
```

Fig. 14. ARIMA model

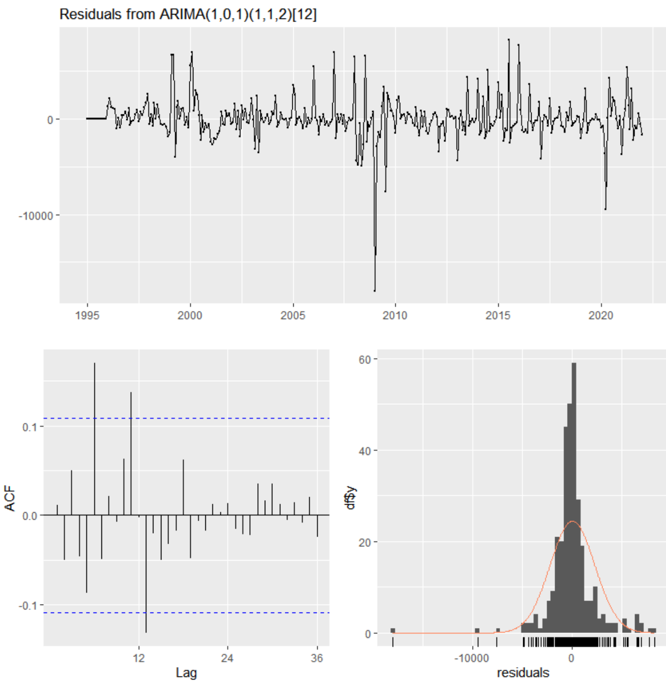


Fig. 15. ARIMA Residuals

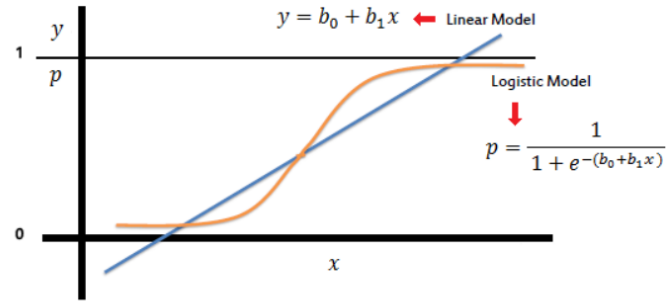


Fig. 16. Logistic Regression Curve

## II. Logistic Regression

### A. *Introduction*

One of the most often used Machine Learning algorithms in the Supervised Learning method is Logistic Regression. It's a method for using a set of independent factors to predict a categorical dependent variable. The output of a categorical dependent variable is evaluated using logistic regression. As a result, there must be a distinct or categorical conclusion. It can be Yes or No, 0 or 1, true or false, and so on, but it always returns probabilistic values rather than precise integers like 0 and 1. Logistic Regression is comparable to Linear Regression in terms of application.

Regression problems are solved using linear regression, whereas classification problems are solved using logistic regression. Instead of fitting a regression line, we fit a "S" shaped logistic function in logistic regression, which predicts two upper and lower boundaries (0 or 1). The curve of the logistic regression function reflects the probability of occurrences such as whether an email is spam or not, if money is good or not, and so on.

### B. *Problem Description*

The purpose of the research is to figure out if a customer has defaulted on a loan based on other characteristics including gender, age, retired, years of schooling, home equity credit, credit card debt, and other debt. When you've finished all of the assumptions and passed all of the model fit tests, you'll have a model that best matches the data.

### C. *Descriptive analysis of Dataset*

A dataset was picked from the financial institution's customer data. Following variables were chosen as dependent and independent variables to meet the objectives.

TABLE I
Data Description

| Variable Name | Data Type | Dependency |
|---|---|---|
| gender | Numeric | Independent |
| age | Numeric | Independent |
| ed | Numeric | Independent |
| retire | Numeric | Independent |
| income | Numeric | Independent |
| creddebt | Numeric | Independent |
| othdebt | Numeric | Independent |
| default | Numeric | Dependent |
| marital | Numeric | Independent |
| homeown | Numeric | Independent |

- *Dependent Variable:*
  1. default : If the person has a loan default on file, the number is 1, otherwise it is 0. The dependent variable consequences should not be associated to use a logistic regression strategy, i.e., in this example, the result is either 1 or 0, where 1 indicates if the consumer has defaulted on their loan and 0 indicates that they have not. There is no such thing as a "middle value" or "output." This clearly demonstrates two distinct but not mutually exclusive outcomes.This criteria is therefore satisfied.

- *Independent Variables:*
  1. gender : indicates the customer's gender. If male, then 0 and if female then 1.
  2. age : provide the customer's age.
  3. ed : Years of Education in Years of a Customer.
  4. retire : 0 if you don't want to retire and 1 if you do.
  5. income : a customer's income in thousands of euros.
  6. creddebt : a customer's credit card debt in thousand euros.
  7. marital status : if married, 1; if not, 0.
  8. homeown : 1 if the customer owns a house, 0 if the

customer is renting.

| | gender | age | ed | retire | income | creddebt | othdebt | default | marital | homeown |
|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 2721.000000 | 2721.000000 | 2721.000000 | 2721.000000 | 2721.000000 | 2721.000000 | 2721.000000 | 2721.000000 | 2721.000000 | 2721.000000 |
| **mean** | 0.517457 | 43.914002 | 14.761852 | 0.113561 | 54.691290 | 2.208151 | 3.929531 | 0.429989 | 0.474825 | 0.634326 |
| **std** | 0.499787 | 17.794929 | 3.270955 | 0.317336 | 60.137589 | 4.334525 | 6.026252 | 0.495165 | 0.499458 | 0.481707 |
| **min** | 0.000000 | 18.000000 | 6.000000 | 0.000000 | 9.000000 | 0.001364 | 0.016704 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 0.000000 | 28.000000 | 12.000000 | 0.000000 | 23.000000 | 0.424116 | 1.053918 | 0.000000 | 0.000000 | 0.000000 |
| **50%** | 1.000000 | 42.000000 | 15.000000 | 0.000000 | 37.000000 | 1.000360 | 2.196096 | 0.000000 | 0.000000 | 1.000000 |
| **75%** | 1.000000 | 58.000000 | 17.000000 | 0.000000 | 64.000000 | 2.272356 | 4.643840 | 1.000000 | 1.000000 | 1.000000 |
| **max** | 1.000000 | 79.000000 | 23.000000 | 1.000000 | 1073.000000 | 109.072596 | 141.459150 | 1.000000 | 1.000000 | 1.000000 |

Fig. 17. Data Description

The Count, Mean, Minimum, 1st Quratiles, 3rd Quratiles, and Maximum values for the data in the Dataframe are shown in fig 17.

### D. Data visualization

Before finishing the model creation steps, we must examine the data visually, such as the connection between variables, scatter diagrams, and histograms. Examine the various factors. Correlation explains the significance of a link between two variables. The correlation values range from -1 to 1, with -1 being the lowest and 1 being the strongest correlation. A negative association is indicated by a -1. (indirect correlation) There is no association if the value is 0. 1 denotes a positive connection (direct correlation). The data co-relation matrix is shown in fig 18.

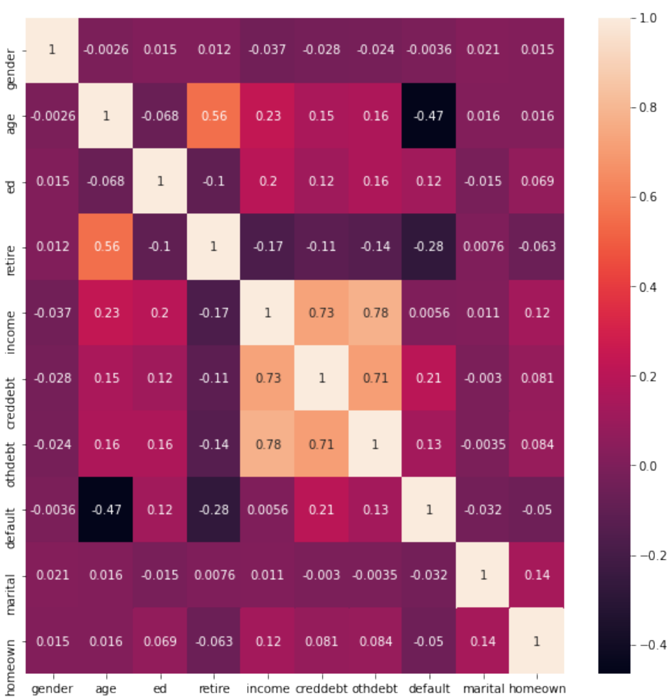


Fig. 18. Data Co-relation

The following is a pair plot of the data (fig 19):

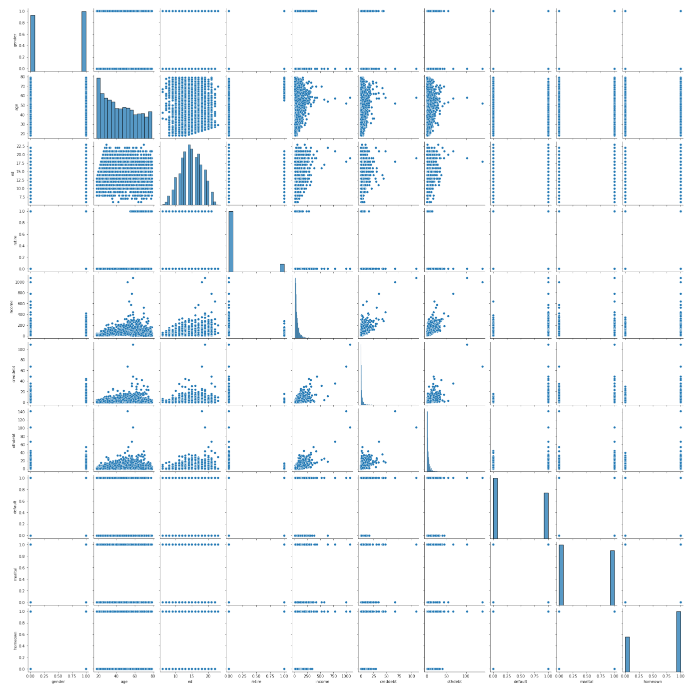

Fig. 19. Pair plot

### E. Model Building

We'll depict the full data set with a box plot since it helps us to quickly notice data set distribution, skewness, and mean values. The box plot below shows that there are several outliers in the variable income. (See Figure 20.)

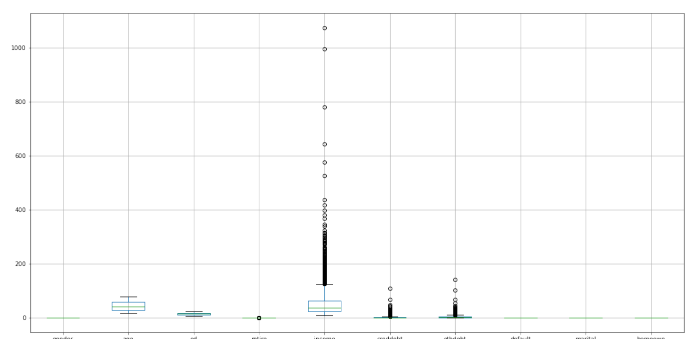


Fig. 20. Data with outlier

The Inter-Quartile Range (IQR) method was used to eliminate outliers, which measures the difference between the data's third and first quartiles. We chose the values of the first and third Quartiles and then used the approach to compute the IQR to acquire the values represented as dots in the graph since we had a number of outliers for the variables 'retire','income','othdebt,' and 'creddebt'. (See Figure 21)

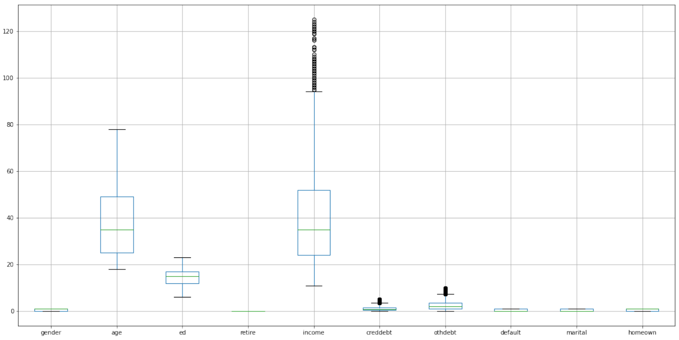


Fig. 21. Data without outlier

After deleting the outliers in fig 22, we used Python's built-in function 'hist' to present the appropriate frequency patterns of each variable.

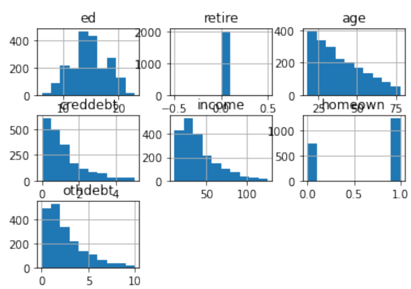


Fig. 22. Histogram after outlier removal

We did statistical OLS regression using statistical function of python function as shown in fig 23.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                default   R-squared:                       0.309
Model:                            OLS   Adj. R-squared:                  0.306
Method:                 Least Squares   F-statistic:                     111.2
Date:                Thu, 05 May 2022   Prob (F-statistic):          1.15e-153
Time:                        21:03:27   Log-Likelihood:                -1075.0
No. Observations:                1999   AIC:                             2168.
Df Residuals:                    1990   BIC:                             2218.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.8430      0.053     15.968      0.000       0.739       0.946
gender        -0.0044      0.019     -0.236      0.813      -0.041       0.032
age           -0.0141      0.001    -19.690      0.000      -0.015      -0.013
ed             0.0133      0.003      4.372      0.000       0.007       0.019
retire     -3.875e-17   6.86e-18     -5.652      0.000   -5.22e-17   -2.53e-17
income        -0.0051      0.001     -9.539      0.000      -0.006      -0.004
creddebt       0.1026      0.010     10.230      0.000       0.083       0.122
othdebt        0.0315      0.006      5.688      0.000       0.021       0.042
marital       -0.0062      0.019     -0.332      0.740      -0.043       0.031
homeown       -0.0509      0.019     -2.610      0.009      -0.089      -0.013
==============================================================================
Omnibus:                      502.475   Durbin-Watson:                   0.628
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               89.584
Skew:                          -0.086   Prob(JB):                     3.53e-20
Kurtosis:                       1.977   Cond. No.                     6.99e+17
==============================================================================
```

Fig. 23. OLS Regression Results

**Logistic Regression Model 1 :**

We choose the variables 'age', 'ed', 'retire', 'income', 'creddebt', 'othdebt', 'marital',and 'homeown' for the first Logistic Regression model. 'gender' was removed because the p-value was 0.813. Data from split-testing and training in 20% and 80%, respectively.

```
x = clean_df[['age','ed','retire','income','creddebt','othdebt','marital','homeown']]
y = clean_df['default']
```

Fig. 24. Logistic Regression Model 1

We discovered that the accuracy of the aforementioned model is 75%, and the confusion matrix is presented below (fig. 25)

```
#confusion matrix and accuracy of model
from sklearn.metrics import plot_confusion_matrix

from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(val_y, y_pred)
print('Confusion Matrix : \n\n',cm)
print('\n')
print('Accuracy Score : ',accuracy_score(val_y, y_pred)*100,'%')
```

```
Confusion Matrix :

 [[165  53]
 [ 47 135]]


Accuracy Score :  75.0 %
```

Fig. 25. Result of Logistic Regression Model 1

**Logistic Regression Model 2 :**

For the second Logistic Regression model, we use the variables 'age', 'ed', 'retire', 'income', 'creddebt', and 'othdebt'. Because the p-values were 0.813, 0.740, and 0.009 for 'gender', 'marital', and 'homeown' were excluded (fig. 26). Data from split-testing and training accounts for 20% and 80% of the total.

```
x1 = clean_df[['age','ed','retire','income','creddebt','othdebt']]
y1 = clean_df['default']
```

Fig. 26. Logistic Regression Model 2

The above model has a 75.25 % accuracy, and the confusion matrix is shown below (fig. 27)

```
#confusion matrix and accuracy of model
from sklearn.metrics import plot_confusion_matrix

from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(val_y1, y_pred1)
print('Confusion Matrix : \n\n',cm)
print('\n')
print('Accuracy Score : ',accuracy_score(val_y1, y_pred1)*100,'%')
```

```
Confusion Matrix :

 [[164  54]
 [ 45 137]]


Accuracy Score :  75.25 %
```

Fig. 27. Result of Logistic Regression Model 2

## III. Conclusion

Finally, after all of this investigation, we reached to the conclusion that ARIMA has the lowest RMSE for time series, i.e. 2247.616. Figure 28 displays the projection from February 2022 to June 2022, taking into account the next 6 months forecast. Figure 29 shows plot of Prediction of next 6 months. With the variables ’age’, ’ed’, ’retire’, ’income’, ’creddebt’, and ’othdebt’, our second model for Logistics regression provides higher accuracy.

```
> forecast(arima.fit1,h=6)
         Point Forecast        Lo 80     Hi 80     Lo 95    Hi 95
Feb 2022       11465.143  8506.1363 14424.149  6939.733 15990.55
Mar 2022       10836.261  7374.4803 14298.041  5541.924 16130.60
Apr 2022        8544.153  4781.5412 12306.765  2789.735 14298.57
May 2022        7433.234  3480.9701 11385.499  1388.768 13477.70
Jun 2022        4813.189   738.2133  8888.164 -1418.948 11045.33
Jul 2022       18950.194 14794.6470 23105.741 12594.833 25305.55
```

Fig. 28. Six Month Prediction

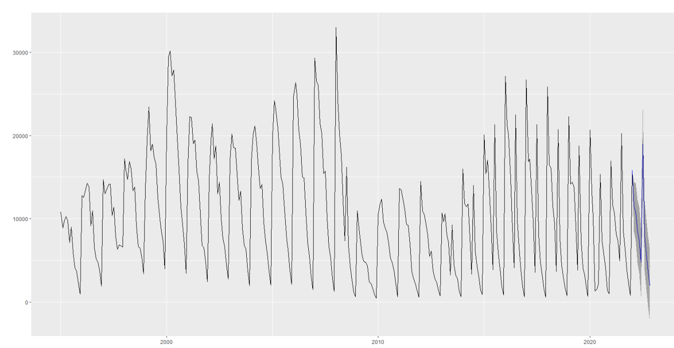


Fig. 29. Six Month Prediction Plot

## References

[1] ”The big data analysis and mining of people’s livelihood appeal based on time series modeling and algorithm”—Liang Lixin ;Lin Lin—2020 International Conference on High Performance Big Data and Intelligent Systems (HPBDIS)—Year: 2020— Conference Paper—Publisher: IEEE

[2] ”Predictive Risk Analysis For Loan Repayment of Credit Card Clients”—Anirudh Bindal;Sandeep Chaurasia—2018 3rd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)—Year: 2018— Conference Paper—Publisher: IEEE

[3] ”Credit Scoring Refinement Using Optimized Logistic Regression”—Hendri Sutrisno;Siana Halim—2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT)—Year: 2017—Conference Paper— Publisher: IEEE