Icon

Description automatically generated

**Conestoga College, Doon Campus**

**Predictive Analytics (1498)**

Statistical Applications for Data Analytics II

PROJECT Report-2: Advanced Data Analytics and Decision-Making Mastery

**Date: 08/14/2024**

**PROJECT BY:**

Prashant Kumar Dua – 8863782

**Guided by:**

Prof. Jonathan Plumbtree

**Contents**

[**Data Exploration 3**](#_Toc174589124)

[**Data Cleaning 4**](#_Toc174589125)

[**Time Series Analysis and Forecasting 5**](#_Toc174589126)

[**Regression Analysis For Forecasting: 9**](#_Toc174589127)

[**Risk Analysis and Monte Carlo Simulation 12**](#_Toc174589128)

[**Linear Optimization and Integer Optimization Models 14**](#_Toc174589129)

[**Non-Linear Optimization 16**](#_Toc174589130)

[**Decision Analysis 17**](#_Toc174589131)

* **Objective:**

The objective of this assignment is to develop a thorough understanding and hands-on experience in using advanced data analytics techniques and decision-making methodologies with Python. Through this assignment, we aim to enhance our skills in analyzing complex datasets, extracting valuable insights, and making informed decisions based on data-driven approaches. By applying these techniques in Python, we hope to strengthen our ability to handle real-world data challenges effectively.

* **Major Covered Topics:**

In this group assignment we covered the following topics for time series analysis and optimization;

1. Time Series Analysis and Forecasting
2. Regression Analysis for Forecasting
3. Risk Analysis and Monte Carlo Simulation
4. Linear Optimization and Integer Optimization Models
5. Non-Linear Optimization
6. Decision Analysis

# **Data Exploration**

For Time series Analysis we use the “ Border Crossing Dataset” from Eastern Border Transportation Coalition (EBTC), a non-profit group of representatives of the constituent state and provincial transportation agencies from Michigan and Ontario.

Data analyzed in this report were collected in Canada. The provinces were responsible for data collection under the direction of Transport Canada.

* **Objective of Analysis:**
* The forecasting approach aimed to create an efficient model replicating existing flows and providing reliable historical trends.
* The U.S. and Canada should collaborate on a streamlined binational process for planning, environmental review, approval, and construction of border crossings.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| **Port name** | **Name of the port of two countries** |
| **State** | **State name** |
| **Port Code** | **Specific code of the port** |
| **Border** | **Border name; Canada-US and Mexico-US** |
| **Date** | **Date of the data** |
| **Measure** | Mode of transportation; bus, train, truck |
| **Value** | No. of vehicles passed |
| **Latitude** | The north-south position of a point |
| **Longitude** | The east-west position of a point |
| **Point** | A specific location on the Earth's surface |

* **Variables Of the Dataset:**

There are 393,000 data with 10 variables which include; Port name, State, Port code, Border, date, Measure, Value, Latitude, Longitude, and Point.

* Python Libraries used in the analysis
* NumPy
* Pandas
* Seaborn
* Matpltlib.pyplot
* Plotly
* Statsmodel.api
* Prophet
* Sklearn.metrics
* Statsmodel.graphic

# **Data Cleaning**

Purpose: Cleaning a new dataset is crucial to have consistent data since data quality directly affects the insights you can draw and the effectiveness of any models you build.

* **Steps for Data Cleaning** :

1. **Handling Missing Values**:

We have used the is.na() to check null values in all the dataset variables. This is important because Null values will create voids in the data and make it difficult to interpret results and provides inaccurate predictions. After running the code in Python, we found no null values in the dataset.

2. **Handling Duplicates**: Duplicate entries add redundancy and can distort statistical analyses, leading to misleading conclusions. Suppose the dataset contains duplicate entries for the same customer’s oil consumption.

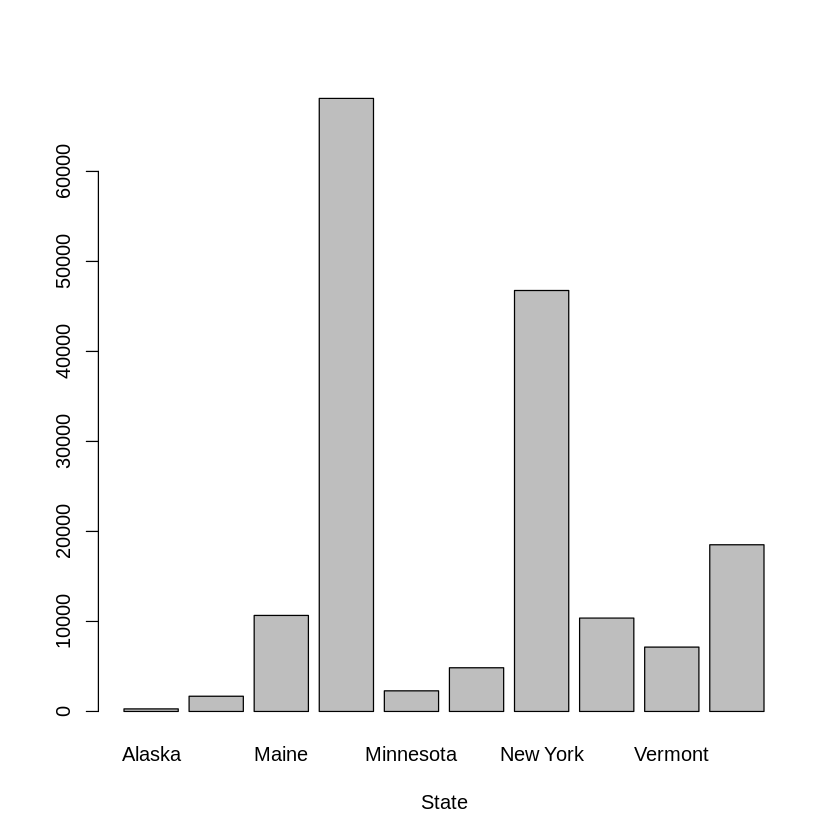
We have used the duplicated () for detecting duplicate entries and after running the code in Python, we found no duplicate values in the dataset.

# **Time Series Analysis and Forecasting**

Time series analysis is a statistical technique used to analyze and model data that is collected or recorded at specific time intervals. It plays a crucial role in understanding patterns, trends, and relationships within time-dependent data.

Time series analysis focuses on the temporal order of data points, recognizing that observations taken over time can exhibit correlations and patterns that are not apparent in cross-sectional data.

Before jumping into the exploratory data analysis; we try to get some insights and patterns of the dataset.



Based on this bar graph, Michigan has the highest number of trucks crossing the Canada-US border. So, for our analysis, we focus on Michigan State.

* **Time series plots and variability:**

A time series plot is a graphical representation of data points in a time series, where time is typically plotted on the horizontal axis (x-axis) and the variable of interest on the vertical axis (y-axis). The plot shows how the data evolves over time, making it easier to observe patterns, trends, seasonal effects, and other temporal relationships within the data.

A green line graph with numbers

Description automatically generated A graph showing a number of trucks crossing

Description automatically generated

From the above time series pattern, we can interpret that there is a constant and static trend from 1996-2009 and we can see seasonal trends before 2009, and after 2010 we can see an upward trend in the trucks.

In the plot above, we try to visualize monthly truck crossing to check in there is more and there is any seasonality exists or not.

**Findings from above plots**:

1. There is an increasing trend in the plot from year 2010 till 2023 except a big dip due to COVID-19 that can be seen in year 2020 due to closing of US-Canada borders at port of entry.
2. Similarly before 2010, we have another significant dip in 2009 due to global recession signifying that impact of
3. Another observation we can is that there are seasonal effects across the time series that should be more evident when we do STL, XTL and classical decomposition.

* **Auto Correlations and lags:**

Autocorrelation is a concept in time series analysis that measures the relationship between a time series' current values and its past values. It helps identify repeating patterns, trends, or dependencies over time.

Positive autocorrelation indicates a persistent trend

Negative autocorrelation suggests a mean-reverting pattern

No autocorrelation suggests randomness or a lack of predictable patterns.

A graph of a truck

Description automatically generated A graph of a truck

Description automatically generated

From the above ACF and PACF Plot, we can see Positive Autocorrelation with rising spikes across the ACF limits showing that the future values of trucks crossing the border have a high correlation with past values.

* **Data Transformation:**

Time series analysis involves data transformation to improve model suitability, stabilize variance, remove trends, convert non-stationary data into stationary, and enhance forecast accuracy.

Here, we perform two different methods for data transformation

1. **Box-Cox Transformation**

A graph with blue lines

Description automatically generated

Here we used the Guererro method for finding the optimal value of lambda that would used for the data transformation. This optimal value is automatically fetched by python functions and fed into the plot.

1. **Logarithmic Transformation**

Next, we used Log transformation to observe how it works on our time series data and below are the results:

A graph of a graph showing the growth of trucks

Description automatically generated

As seen in the above plots the log transformation shows some promising results in capturing the variation within the data while the Box-Cox transformation is not very effective for the same.

* **Time Series Models and Forecasting Techniques:**

Here, we discuss two different forecasting techniques;

1. **Moving Average method:**

A moving average is a calculation to analyze data points by creating a series of averages of different selections of the full data set.

A simple moving average (SMA) is a calculation that takes the arithmetic mean of a given set of prices over a specific number of days in the past.

An exponential moving average (EMA) is a weighted average that gives greater importance to the more recent days, making it an indicator that is more responsive to new information.

1. **Exponential Smoothing:**

Exponential smoothing is used for time series forecasting and helps smooth out irregularities to recognize trends easily. This smoothening of time-series data is helpful for short-term forecasting.

Exponential smoothing works on the principle of weighted average. It means that the values closer to the forecast value are given more significant weightage than older or more distant values.

A graph with blue lines

Description automatically generated

# Regression Analysis For Forecasting:

The regression method of forecasting involves examining the relationship between two different variables, known as the dependent and independent variables. Suppose that you want to forecast future sales for your firm and you've noticed that sales rise or fall, depending on whether the gross domestic product goes up or down.

A graph with blue dots

Description automatically generated A graph of data with blue lines

Description automatically generated with medium confidence

From the above scatter plot, we can conclude that there is a partial positive relationship between value before 2009 and value after 2009.

A barcode diagram with blue dots

Description automatically generated

A comparison of blue and white graphs

Description automatically generated with medium confidence

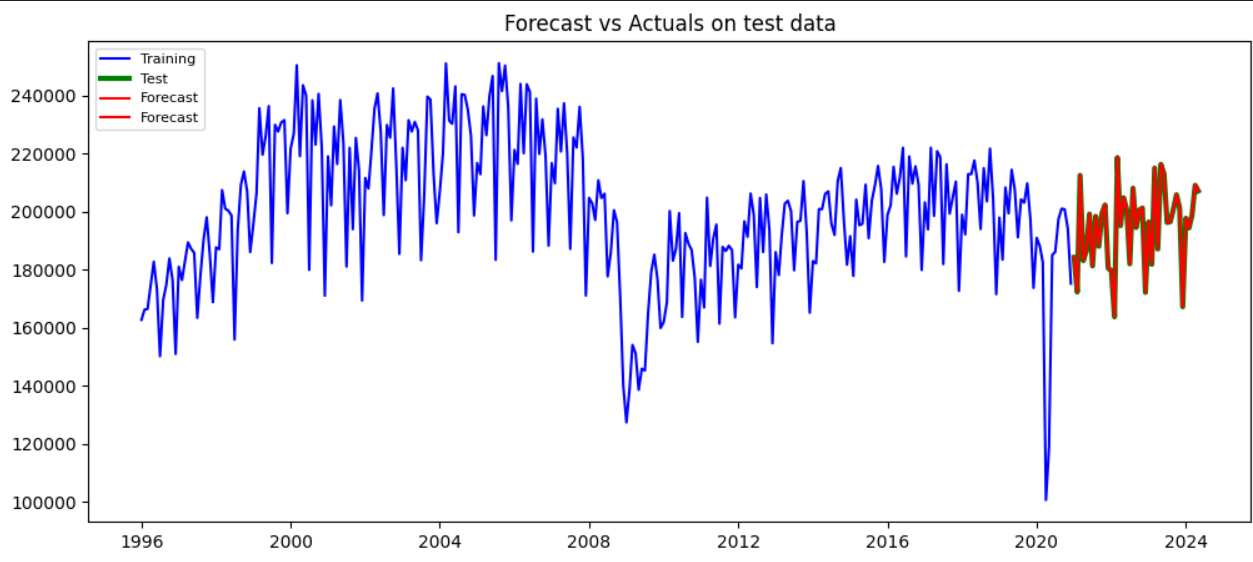
Above Value Vs. Fitted linear forecasting model, there is a positive relationship between Value and fitted.

A graph showing a line of blue and orange lines

Description automatically generated A graph showing the value of a company

Description automatically generated with medium confidence

In the above trend we try to get idea about Original Data Vs. Fitted Trend



A graph with blue lines

Description automatically generated

* **Forecast Accuracy:**

Forecast accuracy is a crucial metric for businesses to assess the reliability and validity of their forecasting models, enabling informed decisions based on predicted outcomes. It compares predicted values with actual observations, indicating areas for improvement.

**Mean Absolute Error(MAE):**

Mean Absolute Error (MAE) is a metric used to measure the accuracy of a predictive model by calculating the average of the absolute differences between predicted and actual values.

**Root Mean Squared Error(RMSE):**

It calculates the square root of the average of the squared differences between predicted and actual values. RMSE gives more weight to larger errors, making it sensitive to outliers. Lower RMSE values indicate better model performance, with a value of 0 representing a perfect fit.

**Mean Absolute Percentage Error(MAPE):**

It measures the average absolute percentage difference between predicted and actual values, expressed as a percentage. Lower MAPE values indicate better model accuracy, with 0% representing a perfect forecast.

* **For our time series models we used MSE and RMSE metrics values as shown below:**

|  |  |  |
| --- | --- | --- |
| **Accuracy** | **Mean Squared Error** | **Root Mean Squared Error** |
| Simple Regression Model | 314951730.84 | 7746.87 |
| Multiple Regression Model | 56385.69 | 237.46 1 |

Hence from above we can see that RMSE values for multiple linear regression model is better because the multiple linear regression uses both trend and seasonal components from decomposition and since our data showed both trend and seasonality, so it makes sense why multiple linear worked in this case.

# **Risk Analysis and Monte Carlo Simulation**

Risk analysis is the process of identifying and analyzing potential future events that may adversely impact a company. A company performs risk analysis to better understand what may occur, the financial implications of that event occurring, and what steps it can take to mitigate or eliminate that risk.

* Risk analysis seeks to identify, measure, and mitigate various risk exposures or hazards facing a business, investment, or project.
* Quantitative risk analysis uses mathematical models and simulations to assign numerical values to risk.
* Qualitative risk analysis relies on a person's subjective judgment to build a theoretical model of risk for a given scenario.
* Risk analysis can include risk-benefit, needs assessment, or root cause analysis.

**Monte Carlo Simulation**

Monte Carlo methods are widely used in various fields of science, engineering, and mathematics, such as physics, chemistry, biology, statistics, artificial intelligence, finance, and cryptography. They have also been applied to social sciences, such as sociology, psychology, and political science.

Monte Carlo methods are mainly used in three distinct problem classes: optimization, numerical integration, and generating draws from a probability distribution.

* **Business Problem:** Wearable Electronic Product Launch

Business sponsors of an electronic company want to find the worst, base, and best scenarios concerning profit earned from selling smartwatches based on the parameters:

**Fixed cost**: $300,000

**Fixed Sales price:** $300 per unit

**Variable Components**:

Demand **varies** between 0 and 20,000 units (most likely $4000)

Variable **labor cost price** varies between $160 and $240 (most likely $200)

We use the Monte Carlo simulation of finding sales metrics as profit with the below assumptions and Earnings metric:

Demand=1000\*np.random.gamma(3,2,1000000)

Variable\_Cost=np.random.uniform(160,240,1000000)

Earnings=(Demand \* Selling\_Price) - (Fixed\_Cost + (Variable\_cost \* Demand))

* We used the Monte Carlo analysis to find all the combinations of variable cost ranges based on:

**Demand shows various gamma distribution values** :

array([5357.63, 3731.49, 8369.91, 4122.61,6395.28, 4156.58])

**Variable cost shows various uniform distribution values** :

array([223.63, 228.03, 224.04, 226.78, 201.66, 162.83])

**Earnings cost shows various values based on the formula**:

array([223.63, 228.03, 224.04, 226.78, 201.66,162.83])

From the Monte Carlo Simulation, we get the following figures.

Maximum Earnings: **4037907.49**

Minimum Earnings: **-295200.62**

Base Earnings: **100000**

Average Profit: **299668.14**

**Conclusion:**

It is profitable for the company to consider the data assumptions as the probability of losses is low and average earnings are quite high

# **Linear Optimization and Integer Optimization Models**

Linear Programming involves optimizing a linear objective function subject to linear equality and inequality constraints. The variables involved in LP models are continuous, meaning they can take any value within a specified range.

**Objective Function:**

The objective function in an optimization problem represents the goal that you are trying to achieve. This could be either maximizing or minimizing some quantity, depending on the problem at hand.

Minimize or Maximize CTX

Where; C is the Cost vector

X is the vector of decision variables

**Constraints:**

Constraints in optimization problem define limitations or requirements for decision variables, restricting feasible regions. They can take forms like inequalities, equalities, or logical constraints.

1. a1​x1 ​+ a2​x2 ​+ ⋯+ an​xn ​≤ b (inequality constraint)
2. a1​x1 ​+ a2​x2 ​+ ⋯+ an​xn ​= b (equality constraint)
3. xi​ ≥ 0 (non-negative constraint)

* **Advertising Budget Allocation**

**Maximization** of Radio advertising and Newspaper advertising audience index based on below parameters:

R – Total budget for Radio

S- Total budget index for Newspaper

**Objective Function**: 50R + 80S

Here, 50 is the mean value of the index for radio and 80 for newspaper advertising

**Constraints**:

1. Total Budget: $1000,
2. Newspaper Budget > = 2\* Radio budget
3. Min newspaper budget: $250
4. Min radio budget: $250

A screenshot of a computer

Description automatically generated

**Interpretation of Optimization Results:**

1. Fun value of -32500 suggest that this the maximum value of audience index based on the optimal value of Radio Sales and Newspaper sales as shown below :

Optimal budget for Radio Sales: $250

Optimal budget for Newspaper Sales: $250

Note that negative sign is shown since python considers this as a minimization problem and hence a negative sign appears just because of that reason.

1. Another important aspect of the optimization results are values under ineqlin which shows results for our inequality constraints:

* **residuals:** It is the difference between the actual values of constraints and the corresponding value obtained from the optimal solution. So, for our problem, we could see that only $500 out of $1000 of the budget is used for an optimal solution which shows some slack is left. Similarly, for the second constraint, $250 is still unused in the constraint indicating how closely constraints are being satisfied at optimal solution.

The residual of 0 in the 3rd and 4th constraints means that the constraint is exactly satisfied at the optimal solution and hence we call these **binding constraints**.

* **Marginals**: also called as shadow price, marginals shows show sensitive the objective function is with increase of decrease of the right side of all the defined constraints. So in our case marginal of -50 for 3rd constraint means that we 1 unit increase in newspaper budget (Assuming all other coefficients in constraints remain constant), audience index reduces by factor of 50 ( remembering that Python considers this as a minimization problem). Similarly , for last constraint a value of -80 shows that with unit increase of radio budget the objective function decreases by 80.

# **Non-Linear Optimization**

Non-linear optimization is a process of finding the best solution in problems where the objective function or the constraints (or both) are non-linear.

* **Objective Function:**

Minimize or Maximize f(x)

where f(x) is a non-linear function of the decision variables x.

* **Constraints:**

Non-linear optimization problems may include both linear and non-linear constraints:

gi(x) ≤ 0 for I = 1,…,m

hj(x) = 0for j = 1,…,p

where gi(x) are the inequality constraints, and hj(x) are the equality constraints.

* **Performance:**

Non-linear optimization techniques can be more computationally intensive, but they are necessary for problems that cannot be simplified to linear forms.

* **Applications:**

Non-linear optimization is crucial in fields like engineering, economics, and finance, where decision-making often involves non-linear relationships. By using Python libraries like SciPy, complex non-linear problems with constraints can be effectively tackled.

A computer screen with white text

Description automatically generated

From the above non-linear optimization, finding a solution where the objective function value is 0.0, with both variables set to 0.0. The process took 5 iterations and required 3 function evaluations and 1 Jacobian evaluation. The gradients at the solution are significant, indicating that the function is sensitive to changes in the variables around this solution.

# **Decision Analysis**

Decision analysis provides a framework for making rational decisions in uncertain conditions by using various techniques like pay-off tables, decision trees, and methods that incorporate probabilities and sample information.

Decision analysis can be used to develop an optimal strategy:

* When a decision maker is faced with several decision alternatives and an uncertain or risk-filled pattern of future events.
* A good decision analysis includes careful consideration of risk.

**Pay-off Tables:**

* Payoff is the outcome resulting from a specific combination of a decision alternative and a state of nature.
* Payoff table is a table showing payoffs for all combinations of decision alternatives and states of nature.

**Decision Tree:**

* A decision tree provides a graphical representation of the decision-making process.

Shows the natural or logical progression that will occur over time.

* Nodes are used to represent decisions and chance events.
* Squares are used to depict decision nodes, circles are used to depict chance nodes.
* **Decision Analysis Techniques**

**Decision Analysis Without probabilities:**

Decision analysis without probabilities is appropriate in situations:

* In which a simple best-case and worst-case analysis is sufficient.
* Where the decision maker has little confidence in his or her ability to assess the probabilities.

There are different types of Approach;

1. Optimistic Approach
2. Conservative Approach
3. Minimax Regret Approach

**Decision Analysis With probabilities:**

There are three different types of approach;

1. **Expected Value Approach**

In decision-making situations where probability assessments for the states of nature are available, we can use the expected value approach to identify the best decision alternative.

The expected value (EV) of a decision alternative is the sum of weighted payoffs for the decision alternative.

1. **Risk Analysis**

Risk analysis helps the decision maker recognize the difference between the expected value of a decision alternative and the payoff that may actually occur.

Decision alternatives and a state of nature combine to generate the payoff associated with a decision.

The risk profile for a decision alternative shows the possible payoffs along with their associated probabilities.

1. **Sensitivity Analysis**

Sensitivity analysis determines how changes in the probabilities for the states of nature or changes in the payoffs affect the recommended decision alternative.

Sensitivity analysis helps the decision maker understand which of these inputs are critical to the choice of the best decision alternative.

**Decision Analysis with Sample information;**

1. **Expected Value of Sample Information**

Decision makers have the ability to collect additional information about the states of nature.

Additional information is obtained through experiments designed to provide sample information about the states of nature.

The preliminary or prior probability assessments for the states of nature that are the best probability values available prior to obtaining additional information.

Posterior probabilities are revised probabilities after obtaining additional information.

1. **Expected Value of Perfect Information**

A special case of gaining additional information related to a decision problem is when the sample information provides perfect information on the states of nature.