AI-DRIVEN EXPLORATION AND PREDICTION OF COMPANY REGISTRATION TRENDS WITH REGISTRAR OF COMPANIES(RoC)

phase2 Project

Propered by:

A.PRIYANKA,

510521205029,

BHARATHIDASAN ENGINEERING COLLEGE,

PHASE2 PROJECT SUBMISSION.

INTRODUCTION

ROC was used initially for radar signal analysis during the world war-II. Currently, ROC analysis is employed in signal detection theory, machine learning, measurement systems, and medical diagnostic applications.

In this article, we will discuss how the ROC analysis can be used for the classification accuracy of computer vision algorithms. The basic concept of ROC analysis can be easily understood with the confusion matrix, also known as the error matrix.

ADVANCED AI ALGORITHM

1. Data Collection: Gather historical data on company registrations from the RoC. This data should include details like company names, registration dates, locations, industry types, and any other relevant information.

2. Data Cleaning and Preprocessing: Clean and preprocess the data to handle missing values, duplicates, and inconsistencies. You may also need to convert text data (e.g., company names) into a structured format for analysis.

3. Feature Engineering: Create relevant features from the data, such as the number of registrations per month, the distribution of registrations by industry, seasonal patterns, and geographical variations.

4. Time Series Analysis: Use time series analysis techniques to identify trends, seasonality, and any underlying patterns in the registration data. This could involve methods like ARIMA, Exponential Smoothing, or more advanced techniques like Prophet or LSTM networks.

5. Machine Learning Models: Develop predictive models using machine learning algorithms. You can use regression, decision trees, random forests, or more advanced models like gradient boosting or neural networks to forecast future registration trends.

6. Natural Language Processing (NLP): If you want to analyze company names or other text data, NLP techniques can be employed. You could use techniques like text classification or sentiment analysis to gain insights from textual information.

7. Geospatial Analysis: If location data is available, geospatial analysis can help you understand regional variations in registration trends.

8. Advanced AI Techniques: Consider advanced techniques like deep learning for time series forecasting or anomaly detection to identify unusual registration patterns.

DEFINITION

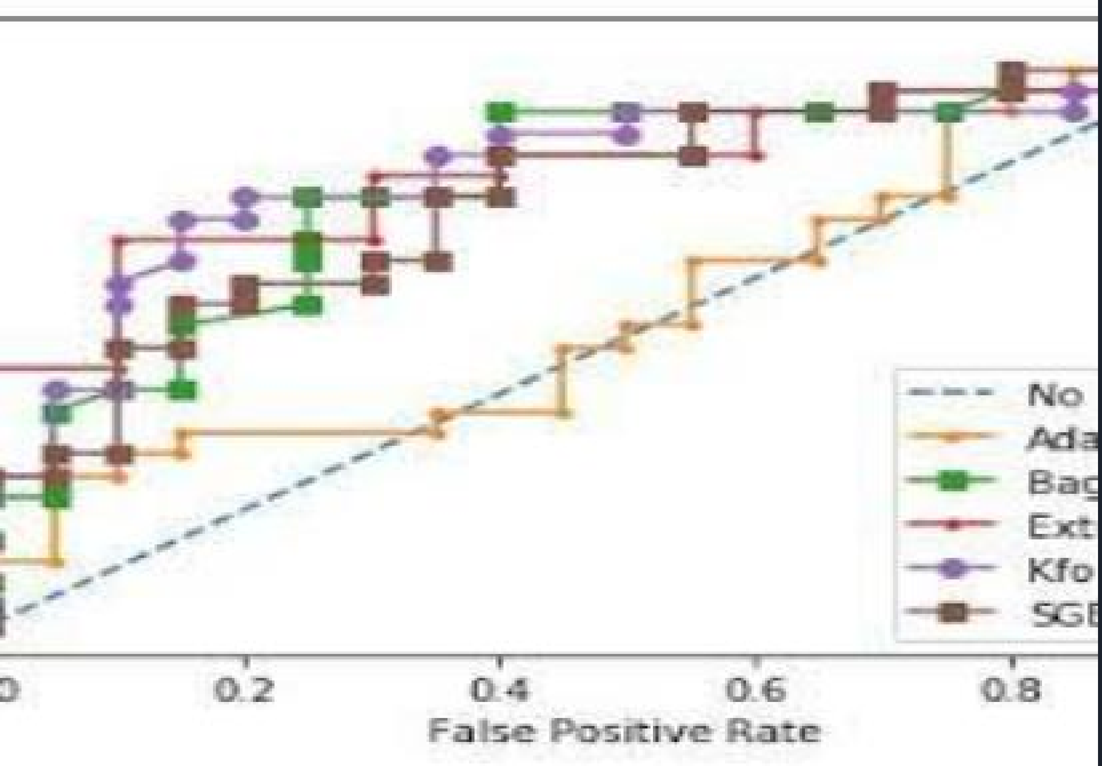
SERIES FORECASTING:

Time series forecasting is a technique for the prediction of events through a sequence of time. The technique is used across many fields of study, from the geology to behavior to economics. The techniques predict future events by analyzing the trends of the past, on the assumption that future trends will hold similar to historical trends.

ENSEMBLE METHOD:

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. To better understand this definition lets take a step back into ultimate goal of machine learning and model building. This is going to make more sense as I dive into specific examples and why Ensemble methods are used.

STEPS FOR ENSEMBLE METHOD



1. Ensemble Technique Selection:

- Choose appropriate ensemble techniques. Common choices include Random Forest, Gradient Boosting (e.g., XGBoost, AdaBoost), and Stacking.

2. Model Training:

- Train multiple ensemble models on the training data. Each model should use a different subset of the data or have different hyperparameters to promote diversity among the models.

3. Model Evaluation:

- Evaluate the ensemble models on the testing data using appropriate metrics like accuracy, F1-score, or ROC-AUC, depending on the nature of the prediction problem.

4. Hyperparameter Tuning:

- Fine-tune the hyperparameters of individual models within the ensemble to optimize their performance.

5. Ensemble Creation:

- Create the ensemble by combining the predictions of the individual models. Common methods include voting (majority or weighted), stacking, or boosting.

6. Performance Evaluation:

- Evaluate the performance of the ensemble model on the testing data and compare it to the performance of individual models.

7. Interpretation and Visualization:

- Interpret the results to gain insights into company registration trends.
- Visualize the trends and predictions using charts, graphs, or dashboards for easy understanding.

8. Deployment:

- If the model meets the desired performance criteria, deploy it for ongoing predictions.
- Set up a pipeline to regularly update the model with new RoC data for continuous monitoring and forecasting.

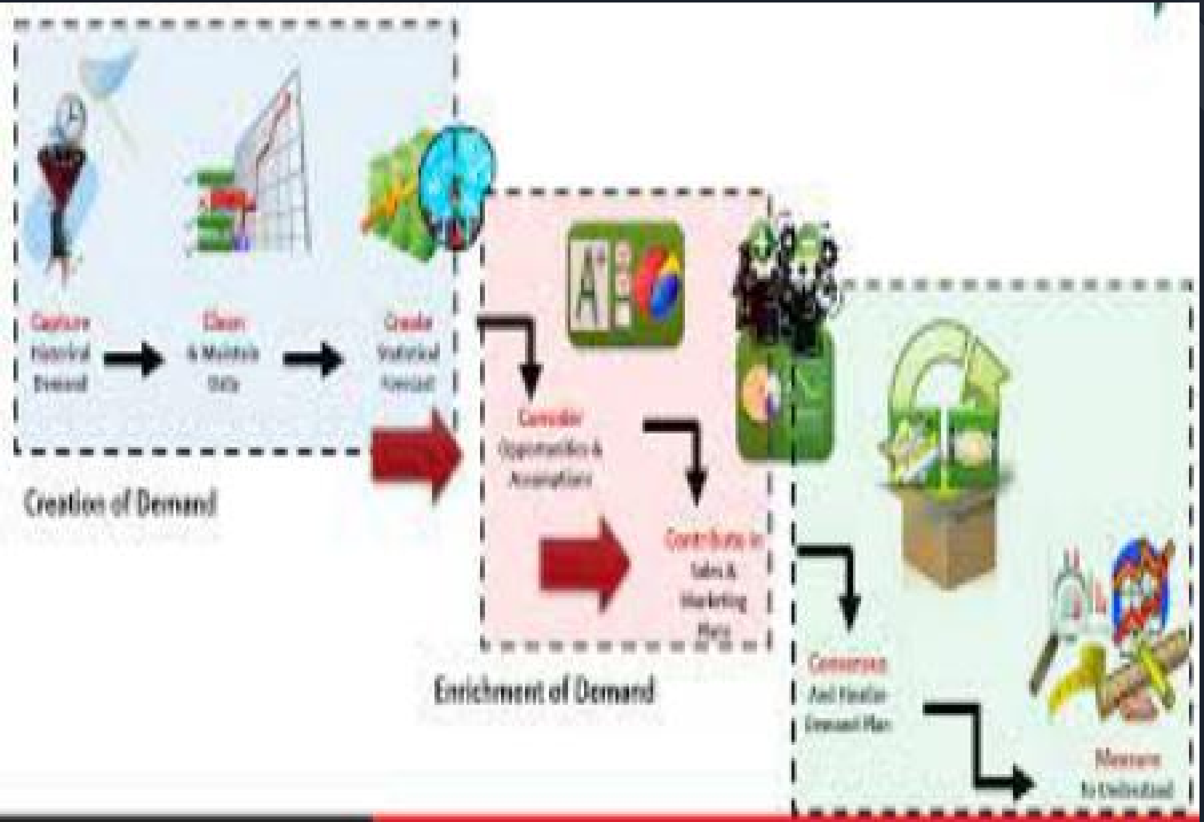
9. Monitoring and Maintenance:

- Continuously monitor the model's performance and retrain it as needed to adapt to changing registration trends.

Remember that the specific implebe taken into account when working

with sensitive company registration data.

STEPS FOR SERIES FORECASTING


Capture Historical Demand
Clean & Maintain Data
Create Statistical Forecast
Creation of Demand
Consider Opportunities & Assumptions
Contribute in Sales & Marketing Plans
Enrichment of Demand
Consensus And Finalize Demand Plan
Measure & Continual

1. Data Collection:

- Gather historical company registration data from the RoC or other reliable sources. This data should include information such as registration dates, types of companies, and geographic locations.

2. Data Preprocessing:

- Clean the data by handling missing values, outliers, and any inconsistencies.
- Convert categorical variables into numerical formats if necessary.
- Ensure the data is in a suitable format for time series analysis.

3. Time Series Analysis:

- Analyze the time series data to identify patterns, trends, and seasonality.
- Use techniques like autocorrelation and partial autocorrelation plots to determine the order of differencing and lag values for forecasting models.

4. Model Selection:

- Choose an appropriate time series forecasting model based on the characteristics of the data. Common models include ARIMA (AutoRegressive Integrated Moving Average), Exponential Smoothing, and Prophet.
- Consider using machine learning models like LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Unit) for more complex patterns.

5. Model Training:

- Split the data into training and validation sets.
- Train the selected model on the training data, tuning hyperparameters as needed.
- Validate the model's performance using the validation set.

6. Forecasting:

- Once the model is trained and validated, use it to make future predictions of company registration trends.
- Generate forecasts for the desired time horizon.

7. Evaluation:

- Assess the accuracy of the forecasts using appropriate metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE).

8. Visualization:

- Visualize the historical data, model predictions, and prediction intervals to communicate the results effectively.

9. Monitoring and Updating:

- Continuously monitor the model's performance and retrain it periodically with new data to ensure its accuracy.

10. Insights and Decision-Making:

- Use the AI-driven predictions and insights to inform business decisions, policy-making, or strategic planning related to company registrations.

SAMPLE PROGRAM

```python
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm

# Load your historical company registration data into a pandas DataFrame
# Replace 'your_data.csv' with the actual file or data source
data = pd.read_csv('your_data.csv')
data['Date'] = pd.to_datetime(data['Date'])
data.set_index('Date', inplace=True)
```

```
# Explore the data with a quick plot
plt.figure(figsize=(12, 6))
plt.plot(data['Registrations'])
plt.title('Company Registrations Over Time')
plt.xlabel('Year')
plt.ylabel('Number of Registrations')
plt.show()
# Time series decomposition to identify trends and seasonality
decomposition = sm.tsa.seasonal_decompose(data['Registrations'],
model='additive')
fig, (ax1, ax2, ax3, ax4) = plt.subplots(4, 1, figsize=(12, 8))
ax1.set_title('Observed')
ax1.plot(decomposition.observed, label='Observed')
ax2.set_title('Trend')
ax2.plot(decomposition.trend, label='Trend')
ax3.set_title('Seasonal')
ax3.plot(decomposition.seasonal, label='Seasonal')
ax4.set_title('Residual')
ax4.plot(decomposition.resid, label='Residual')
plt.tight_layout()
plt.show()
```

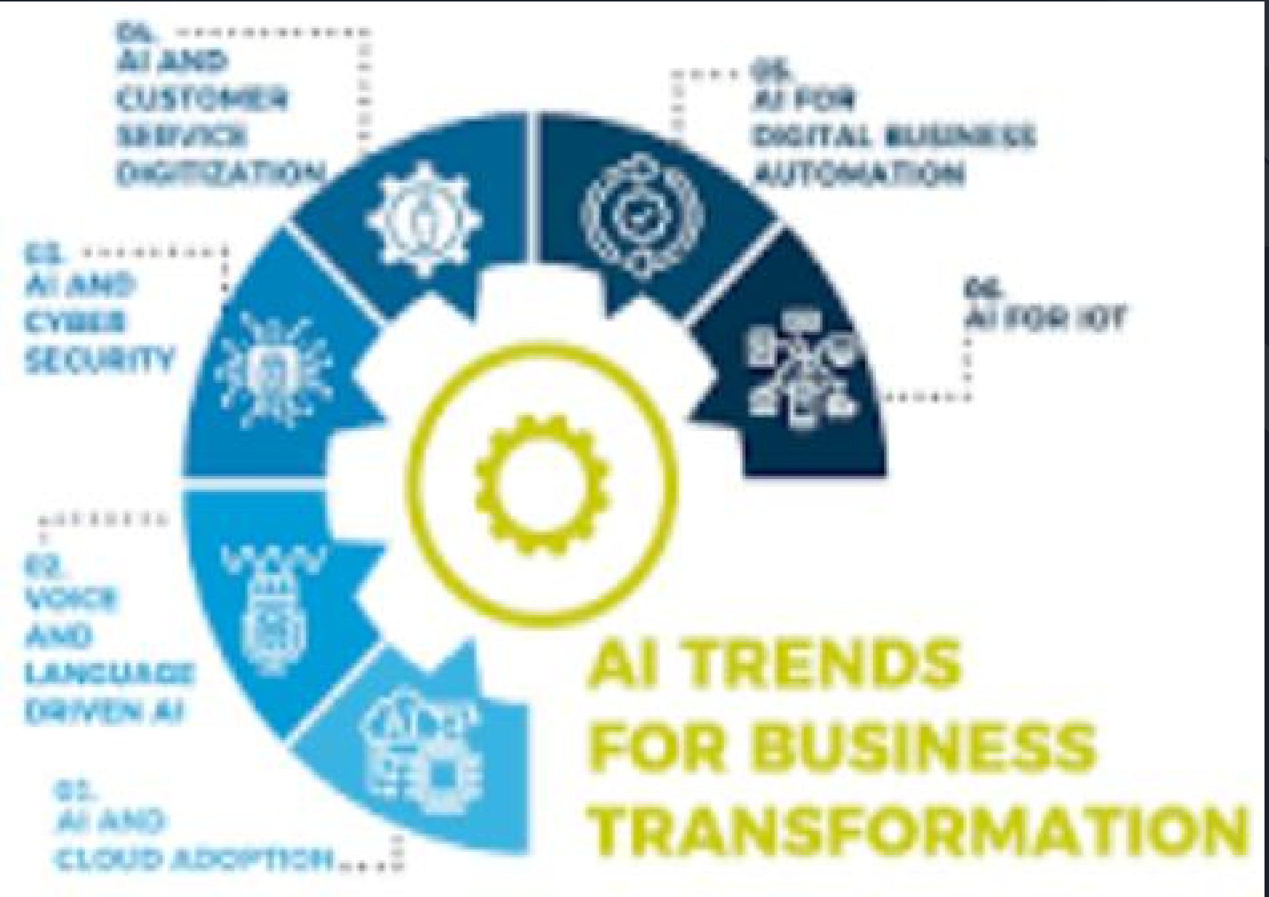
```
# Time series forecasting using SARIMA (Seasonal ARIMA)
# Replace p, d, and q with appropriate values based on your data and analysis
model = sm.tsa.SARIMAX(data['Registrations'], order=(p, d, q), seasonal_order=(P, D, Q, S))
results = model.fit()

# Generate forecasts for future periods (adjust n_periods as needed)
n_periods = 12 # Adjust this for the number of periods you want to forecast
forecast = results.get_forecast(steps=n_periods)

# Extract forecasted values and confidence intervals
forecast_mean = forecast.predicted_mean
forecast_conf_int = forecast.conf_int()
```

~~~
# Plot the forecasted values and confidence intervals
plt.figure(figsize=(12, 6))
plt.plot(data['Registrations'], label='Observed',
color='blue')
plt.plot(forecast_mean.index, forecast_mean.values,
label='Forecast', color='red')
plt.fill_between(forecast_conf_int.index,
forecast_conf_int.iloc[:, 0], forecast_conf_int.iloc[:, 1],
color='pink')
plt.title('Company Registrations Forecast')
plt.xlabel('Year')
plt.ylabel('Number of Registrations')
plt.legend()
plt.show()
```
~~~

OUTPUT:

```
________ Hill Name  Height   Latitude  Longitude
Region
0             Ben Nevis    1345  56.796850  -5.003508
Grampian
1           Ben Macdui    1309  57.070453  -3.668262
Cairngorm
2             Braeriach    1296  57.078628  -3.728024
Cairngorm
3            Cairn Toul    1291  57.054611  -3.710420
Cairngorm
4  Sgòr an Lochain Uaine    1258  57.057999  -3.725416
Cairngorm
```

CONCLUSION


04.
AI AND
CUSTOMER
SERVICE
DIGITIZATION
05.
AI FOR
DIGITAL BUSINESS
AUTOMATION
03.
AI AND
CYBER
SECURITY
06.
AI FOR IOT
02.
VOICE
AND
LANGUAGE
DRIVEN AI
01.
AI AND
CLOUD ADOPTION
AI TRENDS
FOR BUSINESS
TRANSFORMATION

Remember that the quality and availability of data are crucial factors in the success of this endeavor, and the choice of forecasting model should be based on the specific characteristics of the RoC data. Additionally, interpretability and domain expertise are valuable for making meaningful predictions and decisions based on the forecasts.

Thanks!