Project: Election Data

Date of Presentation: 11/05/2024

Name: Nirali Mody

nm17613n@pace.edu

Class Name: Practical Data Science

Program Name: MS in Data Science

Seidenberg School of Computer Science and Information Systems

Pace university

Agenda

- Executive Summary
- Project Plan Recap
- Data
- Exploratory Data Analysis
- Modeling Methods
- Findings
- Recommendations & Technical Next Steps

Executive Summary

This project aims to help **CNBE, a leading news channel**, predict election outcomes through an exit poll model based on voter survey data. By analyzing responses from **1,525 voters** across nine variables—including demographics, political preferences, and economic outlook—the model can forecast the likelihood of voters choosing either the Labour or Conservative party. The goal is to provide CNBE with accurate predictions of each party's overall win and seat coverage, enhancing their election coverage and credibility.

The data in this project represents voter information from the **United Kingdom (UK)**, a country with a rich history of political diversity, where major political parties like the **Labour Party** and the **Conservative Party** hold distinct positions on economic, social, and international issues.

The UK's relationship with Europe has long been a central political issue, notably with Brexit. Voters' attitudes toward Europe provide insights into their broader political views, such as nationalism and trade policy, which may impact party preferences.

Project Plan Recap

Deliverable	Due Date	Status
Data & EDA	11/05/2024	Complete
Methods, Findings & Recommendation	11/12/2024	Complete
Final Presentation	12/03/2024	Not started

Data

Data Details

Variable Name	Description
Vote	Party choice: Conservative or Labour
Age	Age of the voter in years
economic.cond.national	Assessment of current national economic conditions, 1 to 5. (1 - lowest & 5 – highest)
economic.cond.household	Assessment of current household economic conditions, 1 to 5. (1 – lowest & 5 – highest)
Blair	Assessment of the Labour leader, 1 to 5. (1 – lowest & 5 – highest)
Hague	Assessment of the Conservative leader, 1 to 5. (1 – lowest & 5 – highest)
Europe	An 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
political.knowledge	Knowledge of parties 'positions on European integration, 0 to 3.
Gender	Female/Male

- **Data Source:** Election_Data.xlsx file, sourced for analysis to understand voting patterns for CNBE's exit poll project.
- **Sample Size:** 1525 Voters.
- **Variables:** 9 variables, covering demographics, economic assessments, political leader ratings, and political knowledge.
- **Time Period:** Not specified, but assumed recently based on the use for exit poll predictions.

Exploratory Data Analysis

Representation of Votes

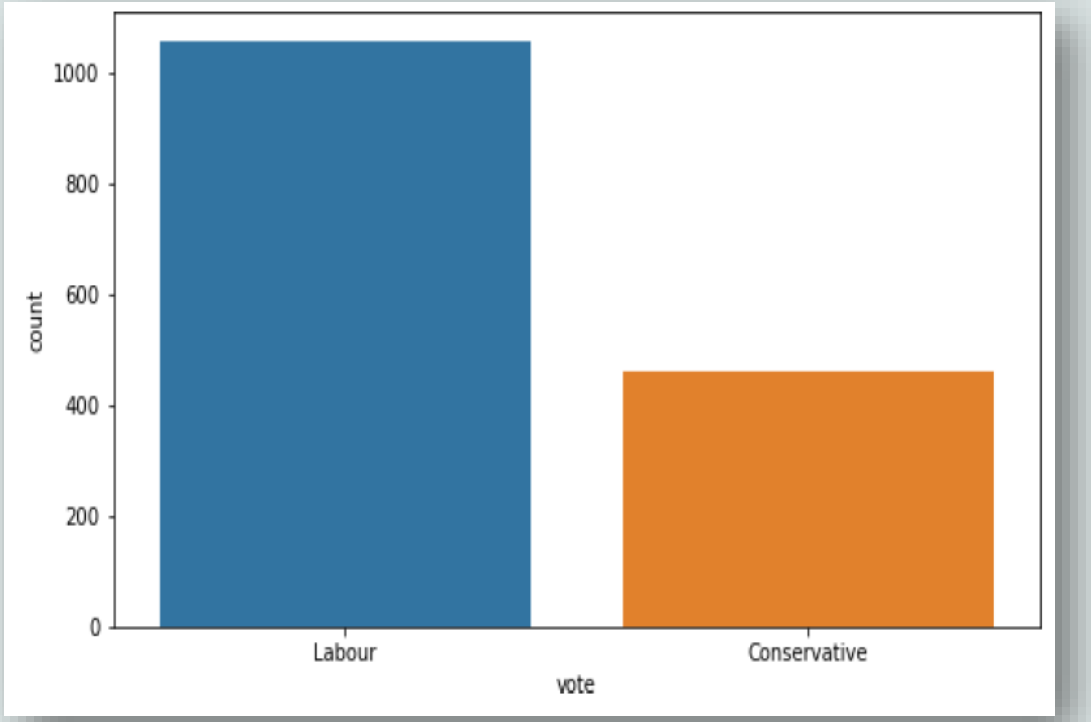
Observation:

- The Labour Party has 1057 votes.
- The Conservative Party has 460 votes.

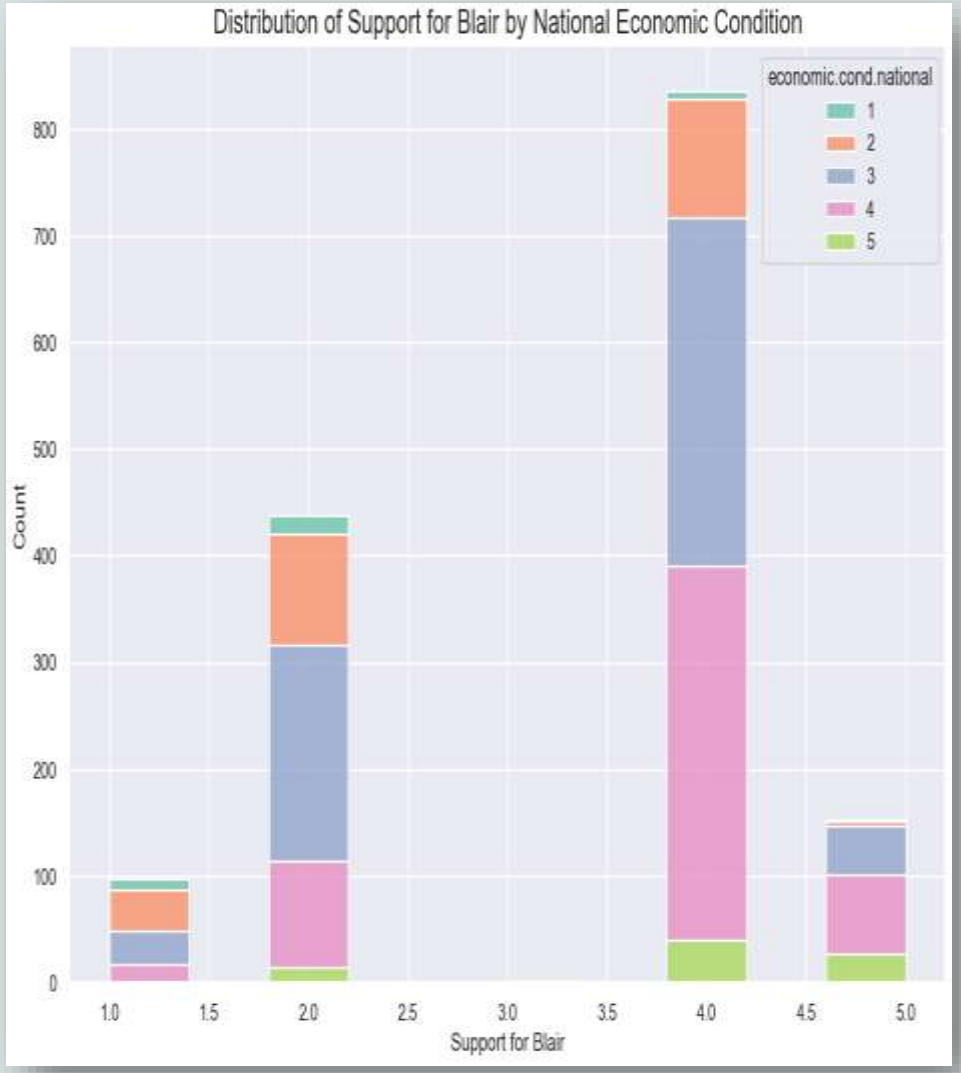
Key Findings:

Labour: If the Labour Party has broader appeal across demographics, it might reflect stronger overall popularity or policies that resonate with a wider range of voters.

Conservative: The Conservative Party may have specific, strong bases among certain demographics. This concentrated support can be an advantage, as targeted messaging to these groups could increase voter turnout and loyalty.



Blair's Popularity & Economic Conditions



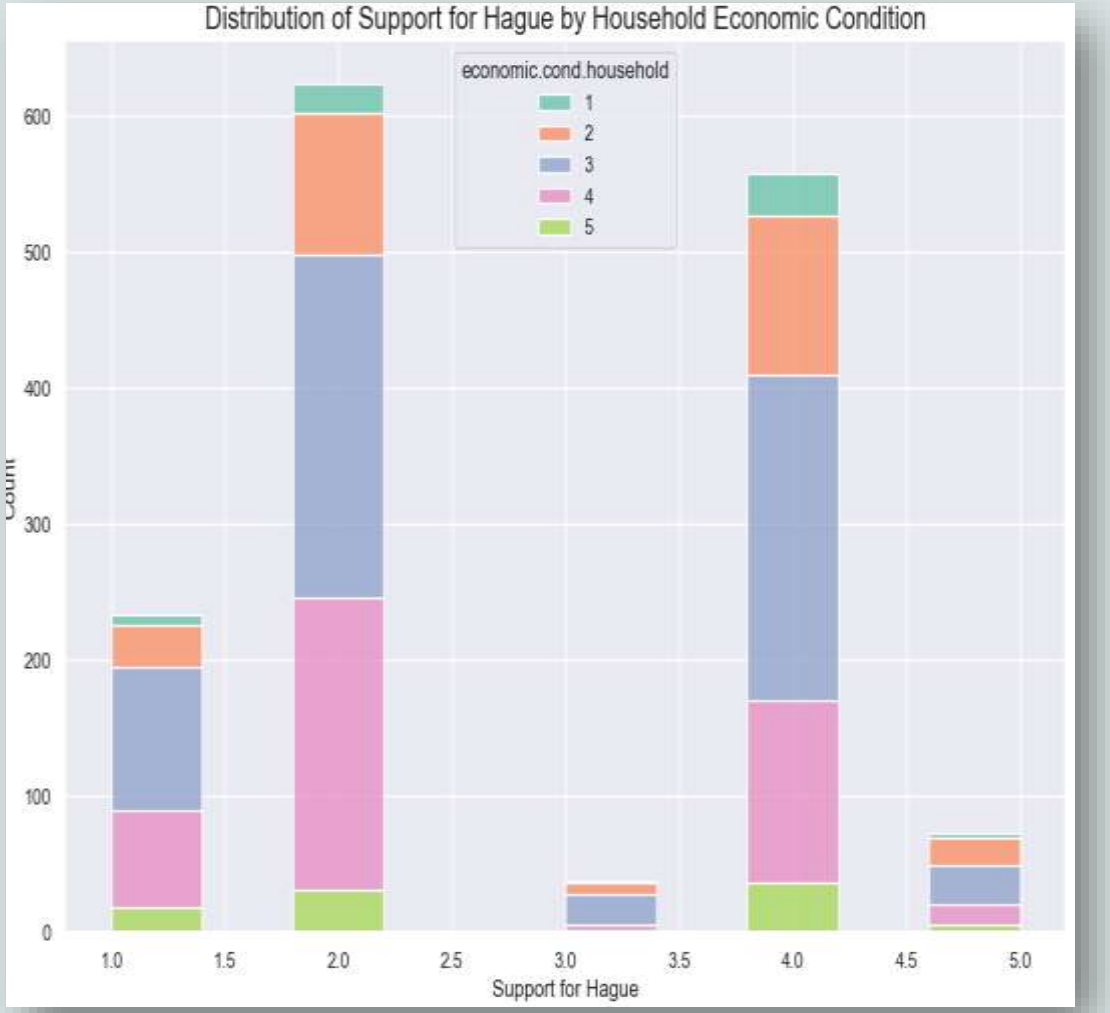
The color with numbers represents Different national economic conditions, likely split into categories such as **"Very poor", "Poor", "Average", "Good", and "Very Good."** Each color represents a different perception of the national economy. The color with numbers represents different levels of household financial satisfaction, possibly categories like **"Struggling," "Stable," "Prosperous," and more.**

What We Looked At: We examined how people's opinions on Blair change with national and household economic conditions—how they feel about the broader economy and their personal finances.

Insights: Blair's support increases when people feel the national economy, or their finances are strong. This suggests that positive economic feelings may increase his appeal to voters.

Possible Actions: If the economy were to improve, Blair's team could highlight this in their messaging, tying his campaign to economic success. Conversely, if economic conditions decline, they might need to focus on alternative strengths, like social issues, to maintain support.

Hague's Popularity & Economic Conditions



Color with numbers represents : Different levels of household financial satisfaction, possibly categories like **"Struggling," "Stable," "Prosperous", and more.** Similar to Blair different levels of household financial satisfaction.

What We Looked At: We also explored how opinions on Hague change with economic conditions to see if there's a similar or different pattern compared to Blair.

Insights: If Hague tends to gain support from voters who feel economically dissatisfied, it might position him as a candidate for change. In contrast, if his support decreases with economic satisfaction, he may appeal more to those who feel disconnected from current economic policies.

Possible Actions: If economic sentiment is low, Hague's team could leverage this by highlighting his policies that aim to improve household finances. If the economy improves, they may want to adjust their approach to maintain relevance.

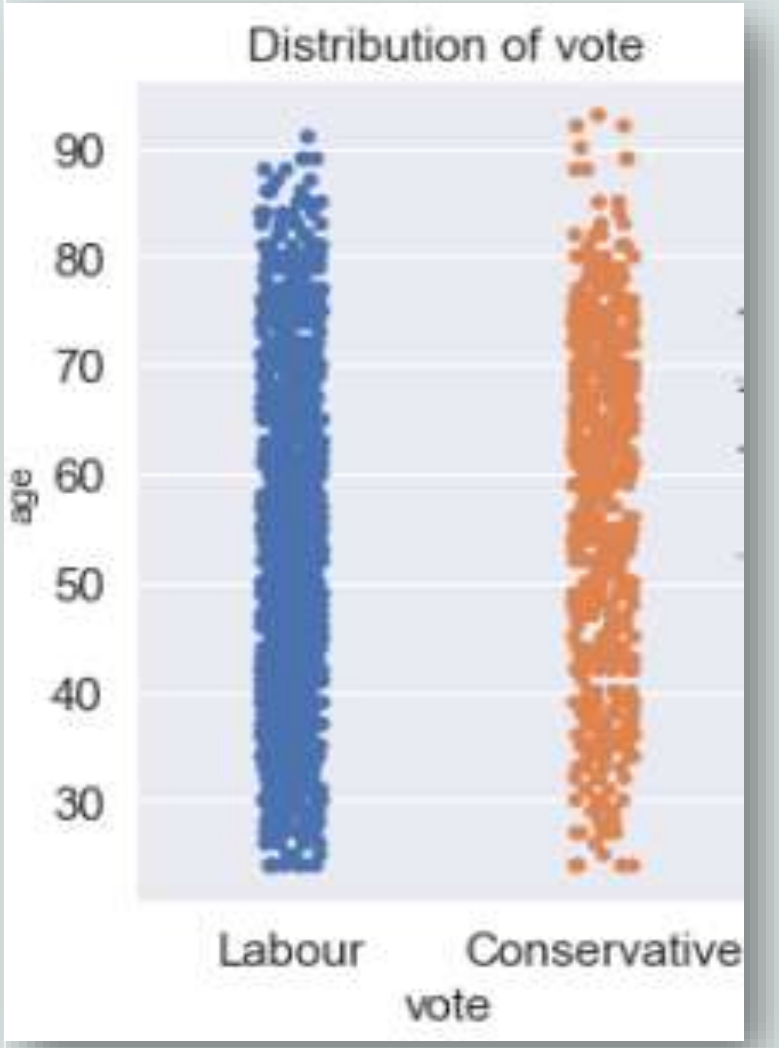
Distribution of Voters' Age

This visualization, a **strip plot**, shows the distribution of voters' age in relation to their voting preference (Labour or Conservative). Each dot represents a voter, positioned by their age and choice, allowing us to see the age spread across each party's supporters.

Insight into Party Support:

- **Dense Clusters**: Areas with many dots indicate age groups with strong support for a particular party. For example, if younger voters show a cluster around Labour, it would suggest that Labour's policies resonate more with the youth.
- **Age-Based Trends**: Identifying any differences in age distribution between the parties can help CNBE inform their viewers on which demographics are driving party support.

This visualization directly supports CNBE's goal of providing a clear and data-backed analysis of voter behavior, adding depth to their election predictions by revealing age-based voting patterns.

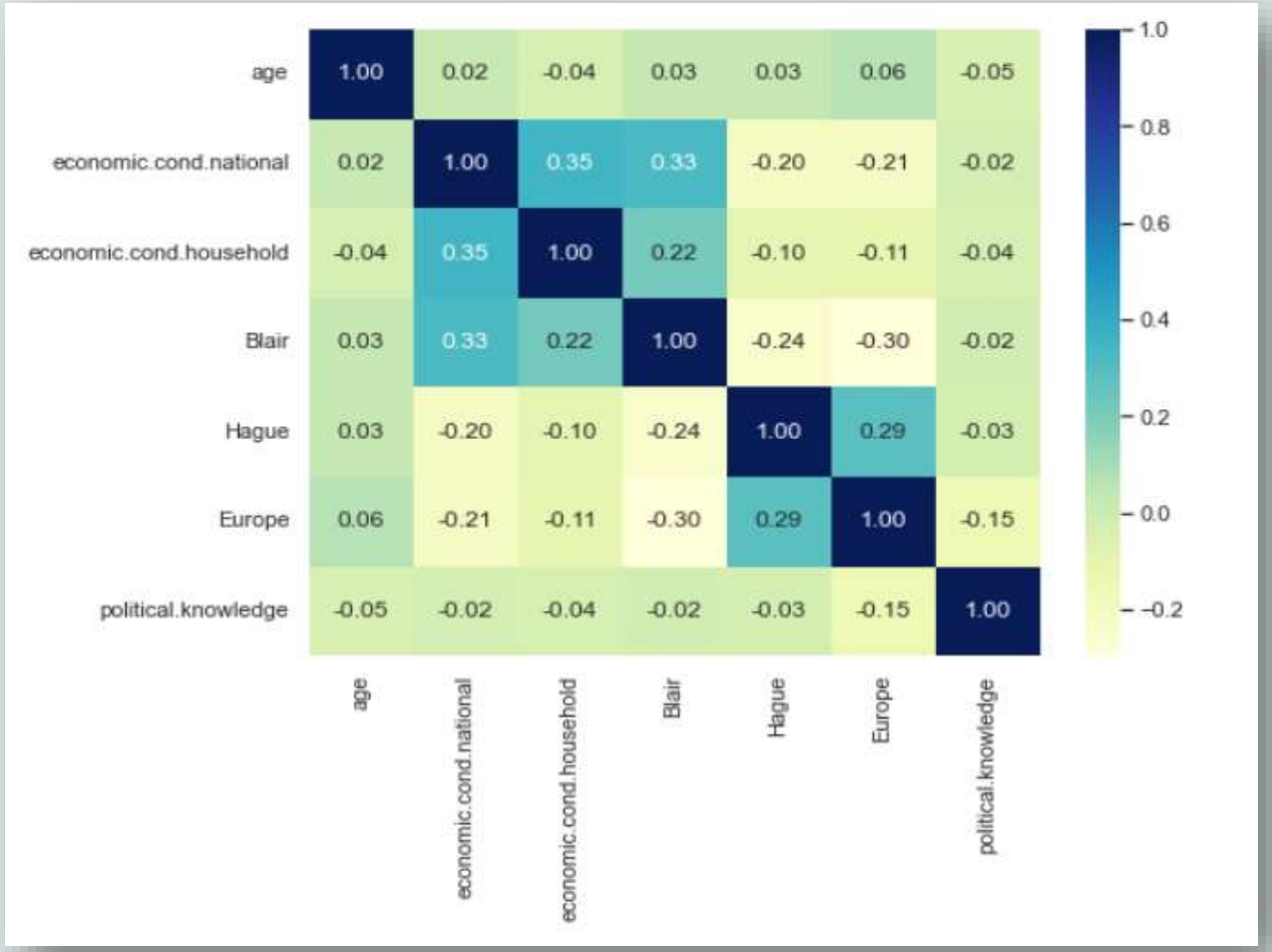


Correlation Heat Map

Key findings from the Correlation Matrix:

- **Economic Conditions (National & Household)**: Strong positive correlation; voters who view national economic conditions favorably tend to feel the same about household finances, potentially influencing party choice.
- **Leader Ratings (Blair vs. Hague)**: No strong correlation; opinions on Labour and Conservative leaders are distinct, showing that voters may favor one leader without necessarily disliking the other.
- **European Integration (Euroscepticism)**: High scores (Eurosceptic sentiment) align more with Conservative support, suggesting that attitudes toward Europe are a key predictor of party preference.
- **Political Knowledge**: Levels of political knowledge correlate with party loyalty, indicating that well-informed voters are more likely to have a strong party preference.

Click here. (For additional information)



Why This Matters for Predicting Election Outcomes:

The correlation matrix helps us identify which factors might work together to shape a voter's preference. For instance:

- Voters with positive views on national and household economic conditions might favor the Labour party.
- High Eurosceptic sentiment could increase the likelihood of a Conservative vote.

Modeling Method

Logistic Regression

The logistic regression model's outcome variable—**voter's party choice**—is the key to making data-driven decisions about how best to reach and engage potential voters.

Why Logistic Regression?

- Logistic regression is especially good at predicting binary outcomes, which means there are two possible answers (e.g., yes/no, support/do not support). Here, it helps us predict if a voter is more likely to support Labour or Conservative based on their characteristics and opinions.

What does it help Predict?

- Likelihood of Support: Based on a voter's profile—such as age, political awareness, economic opinions, or views on leaders—logistic regression estimates the probability that they support either the Labour or Conservative party.

(Refer to this slide for detailed information)

Key Features for Predicting the Model

1. Economic Condition (National)

- **Purpose**: Captures voters' views on the country's economy, often influencing political leanings.
- **Hypothesis**: Voters with negative national economic perceptions are likelier to support Labour, aligning with its focus on economic reform.

2. Blair Favorability

- **Purpose**: Measures receptiveness to Labour's policies through voter sentiment toward Tony Blair.
- **Hypothesis**: Higher favorability towards Blair correlates with Labour support, as it indicates alignment with the party's values.

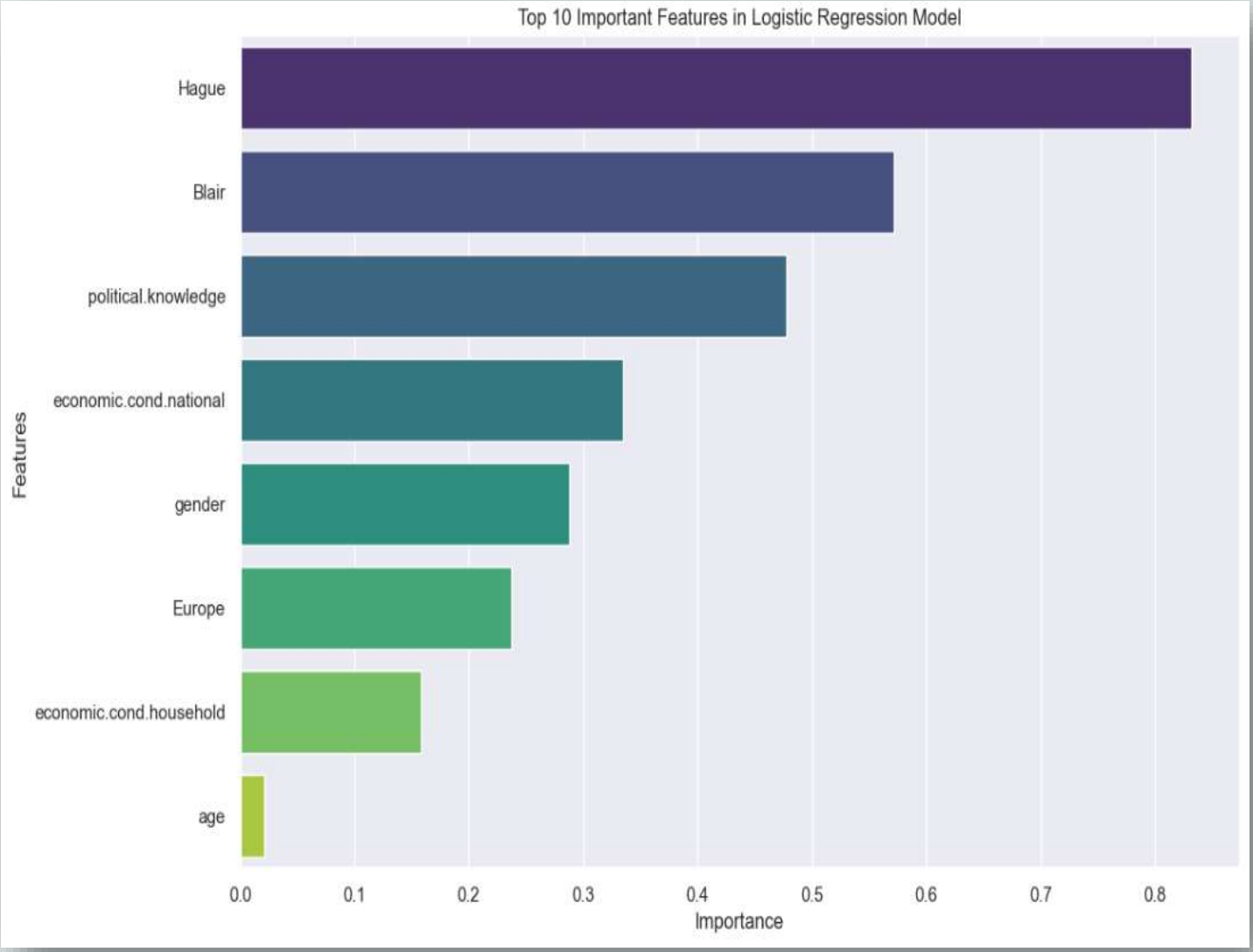
3. Age

- **Purpose**: Reflects generational political trends, as younger voters may favor progressive policies.
- **Hypothesis**: Younger voters are expected to show stronger support for Labour's progressive stance.

- **Initial Assumptions**: Considered broader indicators like *Europe Attitude* and *Household Economic Condition* to capture national and personal perspectives.
- **Refinements**: Analysis showed *National Economic Condition* had a stronger impact on Labour support than *Household Economic Condition*, aligning better with Labour's national reform focus.

Findings

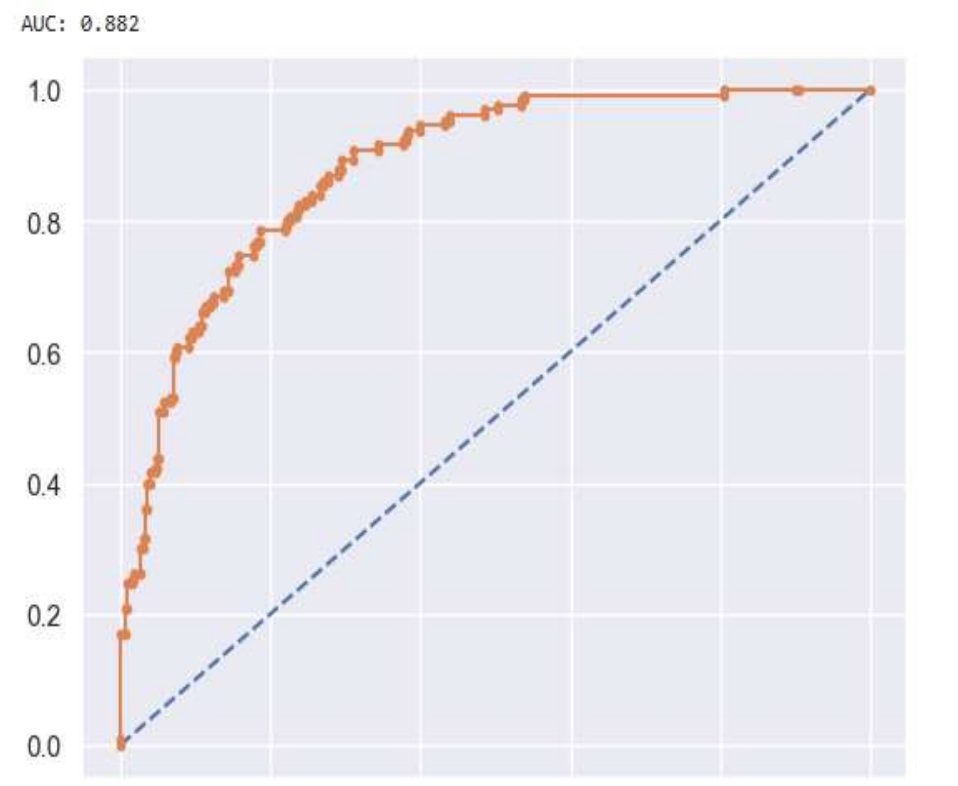
Feature Analysis: Understanding Voter Influences



Leader Perception (Blair, Hague)

- **Relevance:** Voter opinion of party leaders is often a major factor in electoral support, as leaders personify the party's values and vision.
- **Campaign Insight:** Positive or negative sentiment towards leaders can guide campaign strategy, focusing on either promoting the leader's strengths or addressing public concerns. This is crucial for personalizing outreach to resonate with each voter's impression of party leadership.
- (refer to this slide in the appendix for the full list)

Model Performance on Test Data: ROC Curve and AUC Score



ROC Curve & AUC on Test Data

1. What the ROC Curve shows

- The ROC curve is a way to visualize how well the model distinguishes between the two classes—voters likely to support Labour vs. those likely to support Conservative. Each point on this curve represents a different threshold we set to classify voters into one of the two groups. The closer the curve follows the left and top borders of the plot, the better the model is at correctly identifying which group a voter belongs to.

2. The AUC Score: A Measure of Success

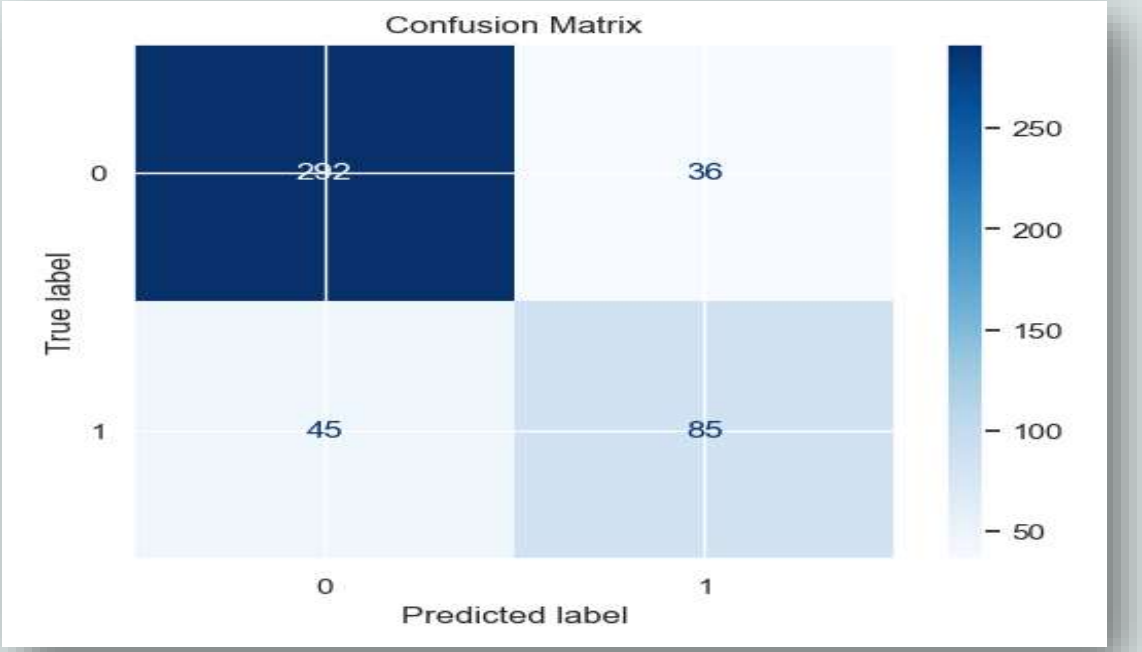
- A higher AUC means better model performance; an AUC closer to 1 would indicate a perfect model, while an AUC of 0.5 would mean the model is no better than random guessing.

(refer to this slide in the appendix for more details.)

Confusion Matrix Insights

In a confusion matrix, each box represents specific outcomes of the predictions made by a model:

- **True Positives (Top-Left):** The model correctly identified 292 instances as Class 0 (likely Conservative voters).
- **True Negatives (85)**: The model correctly identified 85 instances as Class 1 (likely Labour voters).
- **False Positives (36)**: These are instances misclassified as Class 1 (Labour) when they actually belong to Class 0.
- **False Negatives (45)**: These are instances misclassified as Class 0 (Conservative) when they actually belong to Class 1.



(refer to this slide in the appendix for more details.)

Evaluation Metrics (Continuation)

Key Takeaways for Stakeholders

- The model is more accurate in identifying Conservative voters than Labour voters, which could be helpful for targeted messaging.
- Precision is higher than recall for Labour, suggesting that while it is fairly confident in identifying Labour supporters, it may miss some.
- This model provides an 82% accuracy rate, giving a reliable basis for predicting voter affiliation, although further tuning could enhance Labour identification.

Recommendations and Data Science Next Steps

Key Insights and Recommendations

1. National Economic Perception

- **Insights:** Voters' perceptions of a declining national economy strongly predict support for Labour, which aligns with the party's focus on economic reform.
- **Recommendation**: **Emphasize Economic Reform in Messaging**. Focusing on Labour's commitment to national economic improvement can attract voters concerned about the broader economy, especially undecided ones.

2. Blair Favorability

- **Insight**: Positive views of Tony Blair boost Labour support, showing his values still resonate with Labour-aligned voters.
- **Recommendation**: **Utilize Blair's Legacy**. Incorporate Blair's values or legacy into campaign messaging to reinforce Labour's identity and strengthen voter trust and loyalty.

3. Europe Attitude

- **Insight**: Voter sentiment on Europe had minimal impact on party support, suggesting national issues like the economy take priority.
- **Recommendation**: **Focus on National Issues**. Shift campaign resources to emphasize domestic topics like economic reform, healthcare, and education to better align with voter priorities.

Next Steps for Campaign Insights

1. **Consider Advanced Models:**

- Use models like Random Forest or Gradient Boosting to uncover complex relationships.
- Gain deeper insights into how different factors interact and impact voter preferences.
- Identify hidden subgroups within the voter base for more refined targeting.

2. Expand Data Collection:

- Gather data on additional policy areas, such as healthcare and education, to capture broader voter concerns.
- Include more demographic details (e.g., education level, employment status) to create a more complete voter profile.
- Address connected business challenges by understanding voters' top priorities.

Summary for Stakeholders:

To further enhance campaign effectiveness, we recommend exploring advanced modeling techniques for richer insights and expanding our data scope to cover additional voter concerns. This combined approach would strengthen Labour's ability to address the issues most important to voters and optimize campaign resources effectively.

THANK YOU!

APPENDIX

Here's a concise summary for a technical audience:

Project Objective: Analyze voter demographics and preferences to identify key factors influencing support for Labour and Conservative parties, guiding targeted campaign strategies.

Primary Data Insights:

Data Structure: Includes categorical (e.g., vote, gender) and ordinal/numerical data (e.g., economic conditions ratings, age).

Missing Values: None in the dataset; checked for null values and handled duplicates.

Age: Younger voters tend to lean towards Labour, while older voters may favor Conservatives.

Economic Satisfaction: Positive economic views correlate with Labour support; economic dissatisfaction may drive Conservative support.

Political Awareness: Higher awareness levels can affect candidate preference, with nuanced support patterns for Blair and Hague.

Euroscepticism: Strongly correlated with Conservative support, indicating a key voter sentiment.

EDA Techniques:

Distribution Analysis: Histograms and boxplots for demographic breakdowns by party preference.

Correlation Matrix: Identifies relationships between variables, highlighting multi-factor influences (e.g., Euroscepticism and age on party support).

Candidate Support Analysis: Boxplots and strip plots to assess age and awareness impacts on candidate popularity.

Business Relevance:

Insights from EDA guide campaign resource allocation and messaging to resonate with each party's core demographics, maximizing voter engagement potential.

Next Steps: Use insights to refine predictive models for voter support, enabling targeted outreach and optimized campaign strategies.

Why Scaling and Logistic Regression for Predicting Labour Support

1. **Scaling Requirement:**

- *Why It's Needed*: Logistic regression can be sensitive to the scale of features, especially when we have features with different units (e.g., age in years vs. sentiment scores). Scaling, or standardizing, ensures all features contribute equally, preventing larger-scale features from disproportionately influencing the model.
- *Project Relevance*: For instance, a voter's *age* (ranging from 18 to 90+) and *Blair Favorability* (on a smaller, more condensed scale) could otherwise lead to bias in coefficient estimation. Scaling aligns these features, making the model's learning process more stable and consistent across different feature ranges.

2. Why Logistic Regression:

- *Binary Outcome Focus*: The primary task is to predict Labour support, which is a binary outcome. Logistic regression is specifically designed for binary classifications, making it an ideal choice for this yes/no type of question.
- *Probability-Based Output*: Logistic regression provides probabilities for each voter's likelihood of supporting Labour, allowing the campaign team to prioritize outreach by probability scores—targeting voters most likely to be influenced.
- *Interpretability*: Logistic regression estimates how each feature (like *Blair Favorability* or *National Economic Condition*) affects the odds of Labour support. This is valuable because it helps reveal what's driving voter support, allowing us to communicate actionable insights to stakeholders in plain language.

Overall, scaling supports model reliability, while logistic regression delivers interpretable and actionable probability scores that align with the campaign's need for targeted insights into Labour support.

Other Important Features

Demographic Factors (Age, Gender)

- **Relevance**: These fundamental details help identify voter segments based on life stage and gender-related interests or concerns.
- **Campaign Insight**: Understanding which age groups and genders are more likely to support Labour or Conservative can direct messaging to be age- or gender-relevant, like focusing on family policies for younger groups or retirement policies for older ones.

Economic Perception (Economic Condition - National, Economic Condition - Household)

- **Relevance**: These features reflect voters' views on both the national economy and their financial well-being, which are often critical election issues.
- **Campaign Insight**: By identifying which economic concerns resonate with different voter groups, campaigns can highlight relevant economic policies. For example, if the national economic outlook drives support, messaging can focus on the party's plans for economic stability or growth.

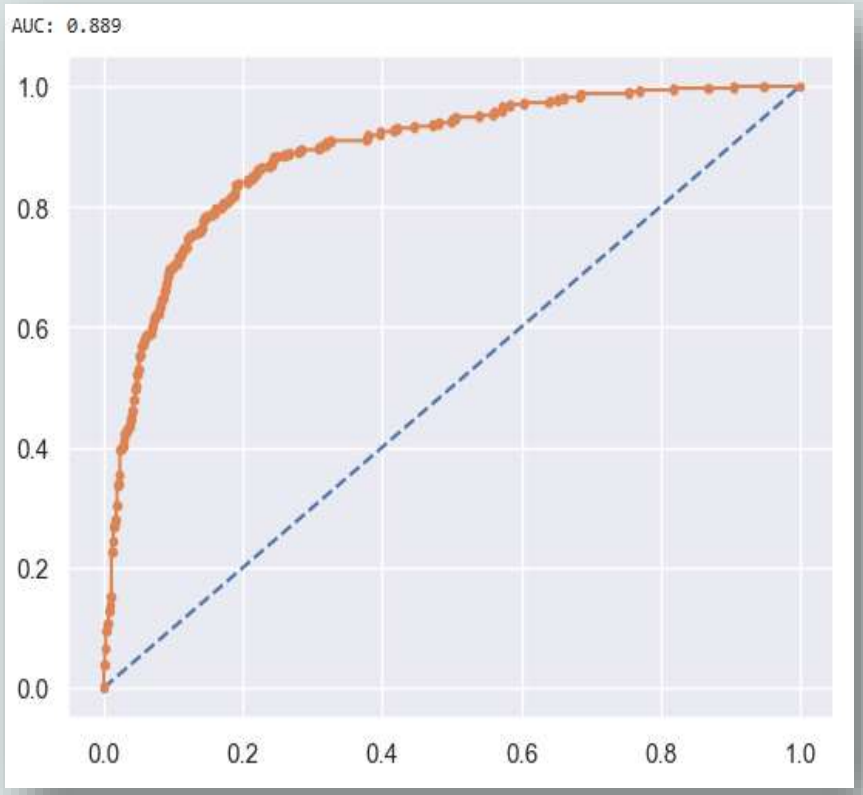
Political Engagement and Awareness (Political Knowledge, Europe)

- **Relevance**: These factors measure a voter's interest in politics and their stance on key issues, such as the UK's relationship with Europe.
- **Campaign Insight**: Understanding the level of engagement and positions on major topics helps the campaign refine its approach. For instance, targeting highly engaged voters with detailed policy discussions on Europe, while using simpler messages for those with less political engagement, can increase overall reach and impact.

Outcome Variable (Vote)

- **Relevance**: This is the actual party preference, either Labour or Conservative, which the model aims to predict.
- **Campaign Insight**: Knowing likely supporters allows for focused, efficient resource allocation, helping the campaign to prioritize high-potential voters and adjust strategies for undecided or opposition-leaning groups.

Model Performance on Train Data: ROC Curve and AUC Score



ROC Curve & AUC on Train Data

1. What the ROC Curve shows

- Our goal is to have a curve that leans towards the top left corner, which would indicate that the model is making correct predictions more frequently than incorrect ones.

2. The AUC Score: A Measure of Success

- In this project, we achieved an AUC score of around (for example) **0.88** on the training data, which indicates that the model is performing quite well in identifying voter preferences.
- An AUC score of 0.88 means that there's an 88% chance that the model will correctly rank a randomly chosen pro-Labour voter higher than a pro-Conservative voter (or vice versa).

3. Why this matters for Campaign Strategy:

- A high AUC score on training data suggests that our model is effective in distinguishing between different types of voters within the data it was trained on.
- This helps campaign strategists identify and focus on specific groups of voters with more confidence, knowing that the model can accurately predict their likely political preference based on the features used (e.g., economic views, age, political awareness).
- However, it's essential to also test the model on new data (like the test data) to ensure it's not just memorizing patterns in the training data but can generalize to unseen voter data, making it a reliable tool for real-world campaign strategies.

Confusion Matrix Insights

Metric	Explanation
Accuracy	Measures the overall correctness of the model in classifying both Labour and Conservative voters.
Precision	Indicates the proportion of voters predicted to support Labour who actually do. Higher precision means fewer false Labour predictions.
Recall	Measures the ability to correctly identify all true Labour supporters. High recall ensures Labour voters are effectively identified.
F1 Score	Balances precision and recall. Useful if the data has class imbalance, as it reflects the model's robustness in handling both types of errors.

Key Takeaways

- **High Accuracy and F1 Score**: The model reliably classifies most voters, indicating it would perform well in real-life scenarios when predicting party support.
- **Balanced Precision & Recall**: Ensures Labour supporters are accurately captured without excessive misclassifications.
- **Implication**: Stakeholders can trust this model's predictions for targeted campaign strategies, especially in distinguishing Labour and Conservative supporters.

Evaluation Metrics (Continuation)

1. Precision:

- **Class 0 (Conservative)**: 0.87 precision, meaning that 87% of the instances predicted as Conservative are correct.
- **Class 1 (Labour)**: 0.70 precision, meaning that 70% of the instances predicted as Labour are correct.
- **Interpretation**: Higher precision for Class 0 indicates that the model is better at correctly identifying Conservative voters without too many misclassifications as Labour.

2. Recall:

- **Class 0**: 0.89 recall, indicating that 89% of actual Conservative voters are correctly identified.
- **Class 1**: 0.65 recall, indicating that 65% of actual Labour voters are correctly identified.
- **Interpretation**: The lower recall for Class 1 means that the model misses some Labour voters, which could be improved for targeting these individuals more accurately.

3. F1-Score

- **Class 0**: 0.88, which balances the model's precision and recall for Conservative voters.
- **Class 1**: 0.68, which balances precision and recall for Labour voters.
- **Interpretation**: The lower F1 score for Labour voters reflects the need for a better balance between precision and recall, as some Labour supporters are missed by the model.