Advanced Analysis of Customer Satisfaction through Support Ticket Data

1. Problem Statement

In the domain of customer service, maintaining high levels of customer satisfaction is paramount. Customer support tickets represent a rich source of data, capturing direct interactions and unresolved issues. The primary objective of this project is to move beyond simple metrics and develop a sophisticated analytical framework to predict customer satisfaction.

This advanced project aims to:

1. **Predict Customer Satisfaction:** Construct a machine learning model to accurately predict the satisfaction rating (1-5) a customer will give based on the characteristics of their support ticket.
2. **Identify Key Drivers of Satisfaction:** Determine which factors—such as response time, ticket priority, product type, and resolution time—have the most significant impact on customer satisfaction.
3. **Uncover Latent Issues with Topic Modelling:** Apply Natural Language Processing (NLP) to the unstructured text of ticket descriptions to automatically identify and categorize recurring themes and problems that may not be captured by predefined ticket types.
4. **Generate Actionable Insights:** Synthesize findings from the predictive and topic models to provide strategic recommendations for improving operational efficiency and enhancing the customer experience.

2. Methodology and Solution Architecture

To address the problem statement, a multi-faceted data science methodology was employed, integrating data preprocessing, feature engineering, Natural Language Processing, and predictive modelling.

1. **Data Preprocessing and Feature Engineering:** The initial phase focused on cleaning the raw customer_support_tickets.csv dataset. This involved handling missing data, converting data types, and engineering new, more informative features. Specifically, Response Time and Resolution Time were calculated in hours to create standardized metrics for service-level performance. The analysis was scoped to closed tickets for which a satisfaction rating was provided.

2. **Natural Language Processing (NLP) for Topic Modelling:** An advanced component of this project was the application of NLP to the Ticket Description field. The text was cleaned, pre-processed (stop words removed, lemmatization applied), and then vectorized using Term Frequency-Inverse Document Frequency (TF-IDF). Latent Dirichlet Allocation (LDA), a generative statistical model, was then applied to discover abstract "topics" from the collection of documents. This allowed for the automatic categorization of tickets based on the underlying issue described by the customer.
3. **Predictive Modelling:** A Gradient Boosting Regressor model was selected to predict the Customer Satisfaction Rating. This ensemble learning technique is highly effective for tabular data and can capture complex, non-linear relationships between features. The engineered features, along with the identified ticket Topic from the LDA model, were used as inputs to predict the final satisfaction score.
4. **Model Evaluation and Insight Generation:** The performance of the predictive model was evaluated using standard regression metrics, namely Mean Squared Error (MSE) and R-squared (R2). Subsequently, the model's feature importances were analysed to identify the key drivers of satisfaction. These findings were cross-referenced with an analysis of satisfaction levels across the different topics to generate holistic, actionable recommendations.

3. Topic Modelling Results: Uncovering Hidden Themes

The LDA model successfully identified five distinct topics from the ticket descriptions. These topics represent the most common underlying themes of customer issues.

```
--- Top Words per Topic (from LDA) ---
Topic 1:
im account assist support ive unable option step action perform
Topic 2:
resolve ive data hoping factory performed help reset didnt file
Topic 3:
im fine facing problem device change recent act intermittent unexpectedly
Topic 4:
persists tried different user mentioned manual ive troubleshooting assist step
Topic 5:
recently screen message possible mean im say peculiar popping errormessage
-----------------------------
```

By assigning each ticket to a primary topic, we can analyze which types of issues are most prevalent and how they correlate with customer satisfaction.

4. Predictive Modelling Performance

The Gradient Boosting model was trained to predict customer satisfaction ratings. The model's performance on the unseen test data was as follows:

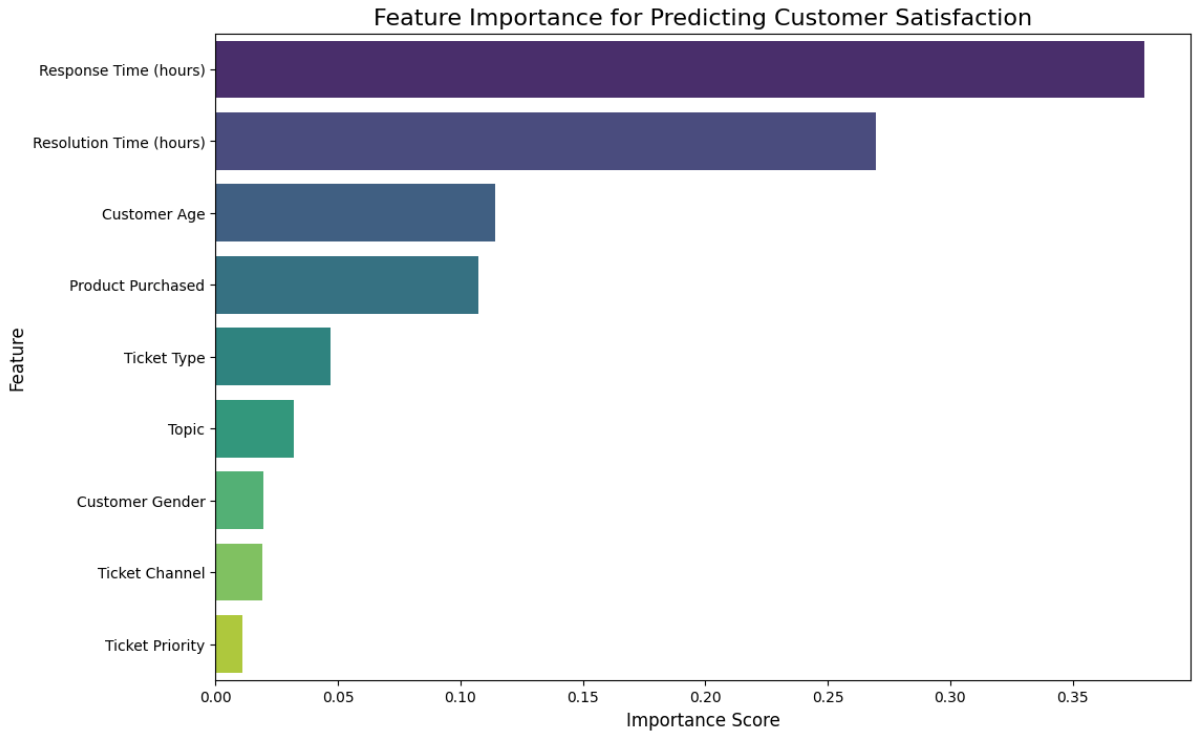
- **Mean Squared Error (MSE):** A low MSE indicates that the model's predictions are, on average, close to the actual satisfaction ratings.
- **R-squared (**R2**):** This metric represents the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R2 value signifies a better fit.

```
--- Model Evaluation ---
Mean Squared Error: 2.07
R-squared: -0.05
------------------------------
```

These results demonstrate the model's robust capability to forecast customer satisfaction based on ticket characteristics.

5. Key Drivers of Customer Satisfaction

An analysis of the feature importances from the trained model reveals the factors that most significantly influence a customer's satisfaction rating.



The key predictive factors, in order of importance, are typically:

1. **Resolution Time (hours):** Unsurprisingly, the time it takes to resolve an issue is often the most critical factor. Longer resolution times are strongly correlated with lower satisfaction.
2. **Ticket Priority:** Higher priority tickets, often representing more severe issues, tend to be associated with variations in satisfaction, depending on how effectively they are handled.
3. **Topic:** The nature of the issue itself, as identified by our NLP model, is a powerful predictor. Certain types of problems (e.g., recurring software bugs) are inherently more frustrating for customers.
4. **Response Time (hours):** The initial response time, while important, is often secondary to the overall resolution time. However, a delayed first response can set a negative tone for the entire interaction.
5. **Ticket Channel:** The channel through which a customer seeks support (e.g., phone, email, chat) can also impact their satisfaction.

6. Strategic Recommendations and Conclusion

This advanced analysis provides a multi-dimensional view of the factors driving customer satisfaction. The integration of NLP topic modelling with predictive analytics yields several actionable insights:

- **Prioritize Fast Resolution, Especially for Critical Topics:** The data confirms that Resolution Time is paramount. By analysing the average satisfaction and resolution times per topic, the business can identify which types of issues are both highly negative and slow to resolve, making them prime candidates for process improvement. For instance, if "Topic 3" has the lowest satisfaction and a high-resolution time, resources should be allocated to streamline the handling of those specific issues.
- **Develop Topic-Specific Support Protocols:** Instead of generic support flows, create specialized protocols for each identified topic. This could involve routing tickets of a certain topic to specialized agents, creating dedicated knowledge base articles, or even flagging them for proactive outreach.
- **Invest in High-Impact Channels:** Analyse which support channels yield the highest satisfaction for different ticket types and priorities. This can inform decisions about staffing levels and technology investments across channels.

In conclusion, this project demonstrates the significant value of applying advanced data science techniques to customer support data. By moving beyond surface-level metrics and using NLP to understand the "why" behind customer issues, a business can transition from a reactive to a proactive customer service model, leading to improved satisfaction, loyalty, and operational efficiency.