# **Part 1: Classical Machine Learning**

# **Objective**

To build a **predictive model for resolution time estimation** using classical machine learning techniques, incorporating both structured and unstructured data (text features).

## **1. Exploratory Data Analysis (EDA)**

### **Overview**

Comprehensive EDA was conducted to understand data distributions, identify outliers, assess correlations, and explore relationships between predictors and target variables.

### **Key Steps**

* **Numerical Analysis:**
  + Summary statistics using .describe() for mean, median, skewness, and standard deviation.
  + Detection of outliers through boxplots and z-score thresholds.
  + Correlation heatmap to understand feature interdependencies.
* **Categorical Analysis:**
  + Frequency counts and visualization of categorical feature distributions.
  + Encoding strategy determined based on cardinality and relevance.
* **Missing Value Treatment:**
  + Imputation for numerical features using median strategy.
  + Mode imputation or “Unknown” category for categorical fields.

### **Insights**

* Certain numerical variables displayed right skewness — addressed through log transformations.
* Key categorical variables like *Issue Type* and *Priority* showed strong associations with resolution time.
* Missing values were mostly concentrated in description-based or optional fields.

## **2. Text Feature Engineering**

Text features were extracted from customer issue descriptions to improve predictive power.

### **Techniques Applied**

| **Feature Type** | **Method** | **Description** |
| --- | --- | --- |
| **TF–IDF Vectors** | TfidfVectorizer | Captured key terms weighted by frequency and document importance |
| **Word Count Features** | CountVectorizer | Represented frequency of words and tokens |
| **Sentiment Scores** | TextBlob | Extracted sentiment polarity and subjectivity values |
| **Text Length Metrics** | Custom Features | Added average word length, total words, and sentence count |

### **Result**

These engineered features enhanced model interpretability and significantly improved performance metrics.

## 

## **3. Model Development and Evaluation**

Initially, a **model benchmarking** phase was performed using the **LazyPredict** library to rapidly evaluate multiple classical machine learning algorithms on the dataset.

This provided a comparative overview of baseline performance across various regressors without manual tuning.From the LazyPredict results, the **top three models** were identified based on *R²* and *RMSE* scores:

* **Ridge Regression**
* **Gradient Boosting Regressor (GBR)**
* **XGBoost Regressor**

These selected models were then **fine-tuned and optimized** using dedicated pipelines, incorporating feature scaling, cross-validation, and hyperparameter tuning.

### **Cross-Validation and Hyperparameter Tuning**

* **5-Fold Cross-Validation** was applied to ensure model generalization.
* **RandomizedSearchCV** was utilized for efficient hyperparameter optimization across parameter grids.
* Models were evaluated using standard regression metrics — *MAE*, *RMSE*, and *R²* — on both training and validation datasets.

## **4. Model Comparison**

### **Evaluation Metrics**

| **Model** |  | **RMSE ↓** | **R² ↑** |
| --- | --- | --- | --- |
| Ridge Regression |  | 0.223 | 0.742 |
| Gradient Boosting Regressor |  | 0.227 | 0.732 |
| XGBoost Regressor |  | 0.230 | 0.726 |

**Best Performing Model:** Ridge Regressor  
 It achieved the lowest RMSE, and the highest R², indicating strong predictive capability.

## **5. Feature Importance and Interpretability**

Feature importance was analyzed for both **Ridge Regression** and **Gradient Boosting Regressor** to identify the key drivers of resolution time.

### 

### **🏆 Top 5 Features — Ridge Regression**

| **Rank** | **Feature** | **Importance** | **Interpretation** |
| --- | --- | --- | --- |
| 1 | **product\_category\_Electronics** | 0.16 | Electronics-related issues tend to require longer handling time. |
| 2 | **customer\_satisfaction** | 0.13 | Lower satisfaction scores correlate with increased resolution time. |
| 3 | **query\_type\_Product Info** | 0.11 | Queries seeking detailed product information take moderately longer. |
| 4 | **product\_category\_Home & Garden** | 0.11 | Home & Garden queries show medium resolution duration. |
| 5 | **customer\_tier\_Gold** | 0.11 | Gold-tier customers often raise more detailed or complex issues. |

### 

### 

### 

### **🏆 Top 5 Features — Gradient Boosting Regressor**

| **Rank** | **Feature** | **Importance** | **Interpretation** |
| --- | --- | --- | --- |
| 1 | **customer\_satisfaction** | 0.27 | Strongest driver; dissatisfaction leads to extended resolution periods. |
| 2 | **query\_type\_Product Info** | 0.12 | Product-related queries dominate issue categories. |
| 3 | **query\_type\_Technical Support** | 0.09 | Technical queries remain among the most time-consuming. |
| 4 | **product\_category\_Electronics** | 0.06 | Electronics consistently appear as high-resolution-time tickets. |
| 5 | **customer\_tier\_Gold** | 0.06 | Gold-tier cases tend to be detailed and require escalation. |

## **6. Business Recommendations (Updated)**

1. **Enhance Satisfaction Monitoring:** The strong impact of *customer\_satisfaction* suggests integrating real-time satisfaction tracking to proactively identify at-risk cases.
2. **Dedicated Product Specialists:** Product information and technical queries account for a significant share of delay; assigning specialized agents to these categories can improve efficiency.
3. **Category-Based Process Optimization:** Electronics and Home & Garden categories repeatedly appear in top features — these should be prioritized for workflow automation or improved documentation.
4. **Tier-Based Prioritization:** Gold and Platinum customers, while valuable, contribute to longer resolution times — improved prioritization and SLA adherence tracking are recommended.
5. **Sentiment & Text Analytics Integration:** Textual patterns (e.g., “keeps,” “weeks”) signal recurring issues; use NLP monitoring to detect and resolve trending concerns faster.

## **7. Deliverables Summary**

| **Deliverable** | **Description** | **Status** |
| --- | --- | --- |
| **Feature Engineering Pipeline** | TF-IDF + Sentiment + Structured Features via ColumnTransformer | ✅ Completed |
| **Model Comparison (MAE, RMSE, R²)** | Evaluation across Ridge, GBR, XGBoost | ✅ Completed |
| **Cross-Validation + Hyperparameter Tuning** | 5-fold CV with RandomizedSearchCV | ✅ Completed |
| **Feature Importance Analysis** | Extracted from tree-based models | ✅ Completed |
| **Business Recommendations** | Actionable insights from model interpretability | ✅ Completed |

## **8. Key Takeaways:**

## Ridge model outperformed ensemble models slightly, indicating linear relationships dominate the dataset.

## Customer satisfaction, product category, and query type are key predictors of resolution time.

## **9. Conclusion**

* The developed classical ML pipeline effectively predicts ticket resolution time by leveraging both structured and unstructured data.
* **Ridge model** achieved the best performance, highlighting the importance of text-derived sentiment and categorical issue features.
* This model provides actionable insights that can help organizations **prioritize, allocate, and manage customer support resources efficiently.**