# Customer Churn Prediction - Logistic Regression Project

## Data Science Test - Data Cleaning and Classification

---

## 1. Introduction

You have been hired as a data scientist at a telecommunications company. Your manager has provided you with a customer dataset and asked you to build a predictive model to identify customers who are likely to churn (leave the company).

**Important:** This is real-world data exported from the company's database, and like most real business data, it contains quality issues that must be addressed before any analysis or modeling can be performed.

---

## 2. Business Objective

Customer churn costs the company millions of dollars annually. Your task is to:

1. Clean and prepare the provided dataset
2. Build a logistic regression model to predict customer churn
3. Evaluate the model's performance
4. Identify the key factors that influence customer churn

---

## 3. Dataset Description

**File:** `customer_churn_raw.csv`
**Size:** Approximately 1,200+ customer records
**Target Variable:** Churn (0 = Customer stayed, 1 = Customer left)

### Features:

| Feature | Type | Description |
|---|---|---|
| CustomerID | Numeric | Unique customer identifier |

| | | |
|---|---|---|
| Age | Numeric | Customer's age in years |
| Gender | Categorical | Customer's gender |
| Tenure | Numeric | Number of months the customer has been with the company |
| MonthlyCharges | Numeric | Amount charged to the customer monthly |
| TotalCharges | Numeric | Total amount charged to the customer over their lifetime |
| ContractType | Categorical | Type of contract (Month-to-Month, One Year, Two Year) |
| InternetService | Categorical | Type of internet service (DSL, Fiber Optic, No) |
| OnlineBackup | Categorical | Whether customer has online backup service (Yes/No) |
| TechSupport | Categorical | Whether customer has tech support service (Yes/No) |
| PaymentMethod | Categorical | How the customer pays their bill |
| PaperlessBilling | Categorical | Whether customer uses paperless billing (Yes/No) |
| Churn | Binary | Target variable - whether customer churned (0=No, 1=Yes) |

---

# 4. Your Tasks

### Task 1: Data Quality Assessment

- Load the dataset and perform an initial exploration
- Identify and document ALL data quality issues present in the dataset
- Create visualizations that illustrate the data quality problems
- Write a summary documentation of the issues found

### Task 2: Data Cleaning and Preprocessing

- Clean the dataset to address all identified issues
- Make and justify decisions about how to handle problematic data
- Ensure the cleaned dataset is ready for machine learning
- Document all transformations and decisions made
- Save the cleaned dataset as `customer_churn_clean.csv`

### Task 3: Exploratory Data Analysis

- Analyze the relationship between features and churn

- Create relevant visualizations to understand patterns in the data
- Identify potential insights about what drives customer churn

### Task 4: Feature Engineering and Preprocessing

- Prepare features for logistic regression (encoding, scaling, etc.)
- Split data into training and testing sets appropriately
- Ensure no data leakage occurs

### Task 5: Model Development

- Implement a logistic regression model
- Train the model on the training data
- Make predictions on the test data

### Task 6: Model Evaluation

- Evaluate the model using appropriate metrics
- Create and interpret a confusion matrix
- Assess the model's performance for the business use case
- Determine which evaluation metric is most important for this business problem and explain why

### Task 7: Interpretation and Recommendations

- Identify which features are most important for predicting churn
- Provide actionable business recommendations based on your findings
- Discuss the limitations of your model

---

## 5. Important Notes

### Expected Ranges (for reference only):

- **Age:** Typical customer age is between 18-70 years
- **Tenure:** Customers typically stay between 1-72 months
- **MonthlyCharges:** Typical monthly bills range from $20-$120
- **TotalCharges:** Should be related to MonthlyCharges and Tenure

### Business Logic (for reference only):

- Customers need internet service to have internet-dependent services
- TotalCharges should logically relate to MonthlyCharges and how long the customer has been with the company

### Evaluation Metrics:

For this business problem, consider which metrics matter most:

- **Accuracy:** Overall correctness
- **Precision:** Of customers predicted to churn, how many actually churned?
- **Recall:** Of customers who actually churned, how many did we identify?
- **F1-Score:** Balance between precision and recall
- **ROC-AUC:** Overall discriminative ability

**Think carefully:** In the business context of customer retention, what is more costly - missing a churning customer or incorrectly flagging a loyal customer?

---

# 6. Deliverables

You must submit:

## 1. Jupyter Notebook (.ipynb) containing:

- Data loading and initial exploration
- Data quality assessment with visualizations
- Data cleaning process (with clear documentation of decisions)
- Exploratory data analysis
- Feature engineering and preprocessing
- Model training
- Model evaluation
- Results interpretation
- Documentation for each step.

## 2. Cleaned Dataset:

- `customer_churn_clean.csv` - Your cleaned and processed dataset

## 3. Presentation:

- **Section 1:** Data quality issues found and how you addressed them (with justifications)
- **Section 2:** Key insights from exploratory analysis
- **Section 3:** Model performance and evaluation
- **Section 4:** Feature importance and interpretation
- **Section 5:** Business recommendations

---

# 7. Submission Guidelines

- Submit all files in a single ZIP folder named: `LastName_FirstName_ChurnProject.zip`
- Ensure your notebook runs from start to finish without errors
- Include all necessary files (notebook, cleaned CSV, Presentation)

---

# 8. Hints (Without Giving Away Solutions)

**General Advice:**

- Start by exploring the data thoroughly before making any changes
- Document your reasoning for every decision
- Test your assumptions about what the data should look like
- Remember that real-world data is messy - that's expected
- There may be multiple valid approaches to cleaning certain issues
- Always verify your cleaned data makes sense before proceeding to modeling

**Questions to Ask Yourself:**

- Are there values that don't make sense for this feature?
- Are there inconsistencies in how the same information is represented?
- Do the relationships between features make logical sense?
- Are there patterns to where data is missing?
- What would be the business impact of different cleaning decisions?

**Common Pitfalls to Avoid:**

- Don't just delete all problematic rows without thinking
- Don't forget to check for logical consistency between related features
- Don't impute missing values before splitting your data
- Don't assume all categorical values are entered consistently
- Don't skip the data validation step after cleaning

---

# 9. FAQ

**Q: How do I know if I've found all the data quality issues?**
A: Systematically examine each feature. Check for missing values, outliers, inconsistent formatting, duplicates, and logical inconsistencies. Create summary statistics and visualizations.

**Q: What if there are multiple ways to handle an issue?**
A: Choose one approach and justify your decision in your documentation. Explain why you chose that method over alternatives.

**Q: How should I handle features with many missing values?**
A: Consider the percentage missing and whether there's a pattern. Document your reasoning for whatever approach you choose.

**Q: Should I create new features?**
A: You may create additional features if you think they'll be useful, but it's not required. Focus first on cleaning the existing features properly.

**Q: What if my model performance isn't great?**
A: Model performance depends on many factors. Focus on doing the cleaning correctly and evaluating thoroughly. Explain any limitations in your presentation.

---

**Good luck! Remember: In data science, cleaning and preparing data properly is just as important as building the model.**