Interim Progress Report (IPR)

Project Title: Customer Churn Analysis in the Telecommunications Industry (IBM)

Module Title: Advanced Computer Science Masters Project (7COM1039-0206-2024)

Student Number & Name: 23037922 – Kinjal Raval

Supervisor Name: Shahrzad Shapoori

1. Background Research and Literature Research

1.1 Introduction

- Customer churn is a significant concern for the telecommunications industry, as retaining customers is more cost-effective than acquiring new ones. This project aims to develop a predictive model to identify potential churners and provide insights to mitigate churn. By leveraging machine learning techniques and data analytics, this study seeks to optimize customer retention strategies for telecom providers.

1.2 Goals and Objectives

- The primary goal of this project is to develop an effective predictive model for customer churn analysis in the telecommunications sector. The specific objectives include:

  Data Collection & Preprocessing: Cleaning and transforming the dataset for effective model training.

  Exploratory Data Analysis (EDA): Understanding patterns and trends in customer behavior.

  Model Development: Implementing multiple machine learning algorithms to compare performance.

  Feature Engineering & Selection: Identifying the most influential factors contributing to customer churn.

  Evaluation & Optimization: Measuring model performance and fine-tuning for optimal results.

  Visualization & Interpretation: Presenting insights in an understandable format for stakeholders.

1.3 Background Research

- Several studies have explored churn prediction using different methodologies. The key findings from the literature review are:

  Traditional Statistical Approaches: Logistic regression has been widely used in churn prediction (Verbeke et al., 2012). While effective for interpretability, it lacks the predictive power of more complex models.

  Machine Learning Models: Random Forest and Support Vector Machines (SVM) have demonstrated higher accuracy in predicting churn due to their ability to capture non-linear relationships (Huang et al., 2019).

  Deep Learning & Neural Networks: Recent research suggests deep learning models outperform traditional models when large datasets are available (Zhao et al., 2021). However, these models are computationally expensive.

Customer Segmentation & Feature Engineering: Studies highlight that customer segmentation based on usage patterns improves prediction performance (Lemmens & Gupta, 2020).

1.4 Practical Research

- To address the challenges in churn prediction, this project incorporates:

  Data Preprocessing Techniques: Handling missing values and normalizing numerical features.

  Model Selection: Testing multiple machine learning models (Decision Trees, Random Forest, XGBoost, etc.).

  Evaluation Metrics: Comparing models using precision, recall, F1-score, and AUC-ROC.

  By integrating theoretical research with practical implementation, this project aims to enhance the accuracy of churn prediction while maintaining interpretability.

---

2. Summary of Progress to Date

2.1 Project Initiation

The project commenced with a **Detailed Project Proposal (DPP)**, submitted on **February 10, 2025**. The proposal outlined key research questions, objectives, methodology, and expected outcomes. Feedback was received, guiding refinements in project scope and methodology.

2.2 Data Preprocessing

The dataset used in this project originates from Kaggle's IBM Telco Customer Churn dataset. The preprocessing steps completed so far include:

- **Handling Missing Data**: Missing values in the TotalCharges column were replaced with zero and converted to numeric.
- **Feature Engineering**: Categorized tenure into groups, transformed categorical variables into numerical values.
- **Data Normalization**: Scaled numerical columns to ensure uniformity across models.

2.3 Exploratory Data Analysis (EDA)

The EDA phase involved analyzing correlations between features and customer churn. Key insights include:

- Customers with month-to-month contracts have a higher churn rate.
- Senior citizens are more likely to churn compared to younger customers.
- Higher charges correlate with increased churn probability.

2.4 Model Development

The project has progressed into the initial stages of model development. Models implemented so far:

- **Logistic Regression**: Provides a baseline accuracy of ~79%.
- **Random Forest**: Improved accuracy but requires hyperparameter tuning.
- **XGBoost**: Shows promising results, balancing accuracy and computational efficiency.

2.5 Next Steps

The upcoming milestones include:

- Further hyperparameter tuning for optimized model performance.
- Enhancing visualization techniques for stakeholder presentation.
- Preparing the **Final Project Report (FPR)** for submission on **April 28, 2025**.

By aligning with project objectives and continuously refining the approach, this project is progressing towards achieving a robust churn prediction model for the telecom industry.

---

3. Ethical, Legal, Professional, and Social Considerations

- **Ethics Approval**: Not required as the dataset is publicly available.
- **Legal Compliance**: Adhering to **GDPR** and data privacy regulations.
- **Bias & Fairness**: Ensuring the model does not disproportionately impact certain customer groups.
- **Professional Standards**: Following best practices in ML development and documentation.

---

4. Project Plan

4.1 Remaining Tasks & Timeline

Task	Start Date	End Date	Dependency
Model Training	Feb 26, 2025	March 4, 2025	Data Preprocessing
Model Evaluation & Selection	March 5, 2025	March 10, 2025	Model Training
Business Insights & Recommendations	March 16, 2025	April 1, 2025	Model Evaluation
Final Report & Submission	April 2, 2025	April 28, 2025	Business Insights

4.2 Evaluation Strategy

- Compare models using **ROC-AUC, confusion matrix, and SHAP interpretability**.
- Perform **hyperparameter tuning** for improved performance.
- Validate findings through **business-oriented case studies**.

---

5. Level of the Project

5.1 Complexity and Depth

- The project integrates both **theoretical research and practical implementation**, demonstrating MSc-level depth.
- Advanced techniques like **SHAP, feature selection, and ensemble learning** are used.

5.2 Testing & Validation

- Conducting **k-fold cross-validation** to ensure model reliability.
- Evaluating **edge cases** and **performance under different conditions**.

5.3 Tools & Justification

- **Python**
- **Pandas & Matplotlib** for data exploration.
- **Jupyter Notebooks** for code reproducibility.

---

6. Referencing and In-Text Citation

- Harvard referencing style is followed.
- Key sources include academic papers on churn prediction and ML methodologies.

Example References:

1. Fader, P. S., & Hardie, B. G. (2013). How to Project Customer Retention. *Journal of Marketing Research, 50(2), 263-280.*
2. Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research, 43(2), 204-211.*
3. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpretable Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS).*

Appendices

Appendix 1: Data Preprocessing Code

(Sample Python code snippets)

```
df_columns = df.columns.tolist()
for column in df_columns:
    print(f"{column} unique values : {df[column].unique()}")
```

```
customerID unique values : ['7590-VHVEG' '5575-GNVDE' '3668-QPYBK' ... '4801-JZAZL' '8361-LTMKD'
 '3186-AJIEK']
gender unique values : ['Female' 'Male']
SeniorCitizen unique values : [0 1]
Partner unique values : ['Yes' 'No']
Dependents unique values : ['No' 'Yes']
tenure unique values : [ 1 34  2 45  8 22 10 28 62 13 16 58 49 25 69 52 71 21 12 30 47 72 17 27
  5 46 11 70 63 43 15 60 18 66  9  3 31 50 64 56  7 42 35 48 29 65 38 68
 32 55 37 36 41  6  4 33 67 23 57 61 14 20 53 40 59 24 44 19 54 51 26  0
 39]
PhoneService unique values : ['No' 'Yes']
MultipleLines unique values : ['No phone service' 'No' 'Yes']
InternetService unique values : ['DSL' 'Fiber optic' 'No']
OnlineSecurity unique values : ['No' 'Yes' 'No internet service']
OnlineBackup unique values : ['Yes' 'No' 'No internet service']
DeviceProtection unique values : ['No' 'Yes' 'No internet service']
TechSupport unique values : ['No' 'Yes' 'No internet service']
StreamingTV unique values : ['No' 'Yes' 'No internet service']
StreamingMovies unique values : ['No' 'Yes' 'No internet service']
Contract unique values : ['Month-to-month' 'One year' 'Two year']
PaperlessBilling unique values : ['Yes' 'No']
PaymentMethod unique values : ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
MonthlyCharges unique values : [29.85 56.95 53.85 ... 63.1  44.2  78.7 ]
TotalCharges unique values : ['29.85' '1889.5' '108.15' ... '346.45' '306.6' '6844.5']
Churn unique values : ['No' 'Yes']
```

[12]:
```
df.describe()
```

[12]:

	SeniorCitizen	tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	30.090047
min	0.000000	0.000000	18.250000
25%	0.000000	9.000000	35.500000
50%	0.000000	29.000000	70.350000
75%	0.000000	55.000000	89.850000
max	1.000000	72.000000	118.750000

[13]:
```
for elem in df.columns[0:]:
    print(df[elem].value_counts())
    print("\n")
```

```
SeniorCitizen
0    5901
1    1142
Name: count, dtype: int64


Partner
No     3641
Yes    3402
```

Appendix 2: Gantt Chart