1. Introduction

Customer churn refers to the process where customers stop their relationship with a business. This is important for businesses because it affects how much money they make and how much they grow. To handle churn, companies need to understand why customers are leaving and come up with ways to keep them happy so they stay or find new customers to replace the ones who left.
Lu's Communications, a well-known telecommunication company in the UK that provides internet, landline, and TV services, is facing the same problem. They are trying to predict customer churn beforehand and They need accurate predictions because if they're wrong, lots of resources might be wasted. The task is to make a report that explains churn prediction to the managers at Lu's Communications. The report should be easy to understand and help them make smart choices. The dataset at our disposal comprises a substantial 7350 observations, featuring attributes such as customer ID, gender, location, partnership status, dependents, seniority, tenure, monthly costs, service packages, customer service scores, and the critical churn classification.
The dataset can exhibit various issues, such as missing values, outliers, or inconsistencies, requiring comprehensive data preprocessing. During exploratory data analysis (EDA), one must assess data distributions, relationships, and patterns, using visualizations and summary statistics to unveil insights. EDA aids in understanding data quality, uncovering biases, and guiding preprocessing decisions. Subsequently, data analytics involves applying statistical and machine learning techniques to extract meaningful information, validate hypotheses, and make informed decisions, addressing challenges like overfitting or imbalanced classes. Overall, these stages collectively aim to transform raw data into a reliable, interpretable, and actionable form, enabling accurate and insightful analysis.

2. Data preparation

The dataset, comprising 7350 observations, was processed to prepare it for algorithmic application and modeling. The features included customer-id, gender, location, partner, dependents, senior status, tenure, monthly-cost, package, survey score, and churn class. Notably, the 'dependents' field contained numerous 'Unknown' values, 'Tenure' exhibited negative or irrational entries, and 'survey' contained 'No reply' entries. Further, gender values were encoded as 1 for male and 0 for female, location values were numerically labeled from 1 to 17 and churn values were also encoded as 1 for churn=no and 0 for churn=1. These issues were addressed by dropping problematic entries, resulting in a refined dataset comprising 4608 observations, ready for subsequent analysis and modeling.

3. Exploratory data analysis and Data pre-processing

A substantial portion of the data was removed during the Data Preparation phase, especially due to the considerable number of 'Unknown' values (2277 instances) present in the 'dependents' feature. Given the binary nature of this feature and the prevalence of 'Unknown' values, conventional imputation strategies such as replacing them with a third category or the median were unsuitable. Consequently, a decision was made to employ classification models to predict the 'Unknown' values.

To determine the most relevant features for this prediction, a correlation matrix was used,Figure 1, revealing that ’partner,’ ’senior,’ and ’Tenure’ exhibited the highest correlations with ’dependents.’ Subsequently, RandomForestClassifier, SVC, and GradientBoostingClassifier were employed for prediction, with all three models achieving an accuracy of 0.76. Among these, SVC was selected (randomly) as the preferred model for imputing the ’Unknown’ values within the ’dependents’ feature.

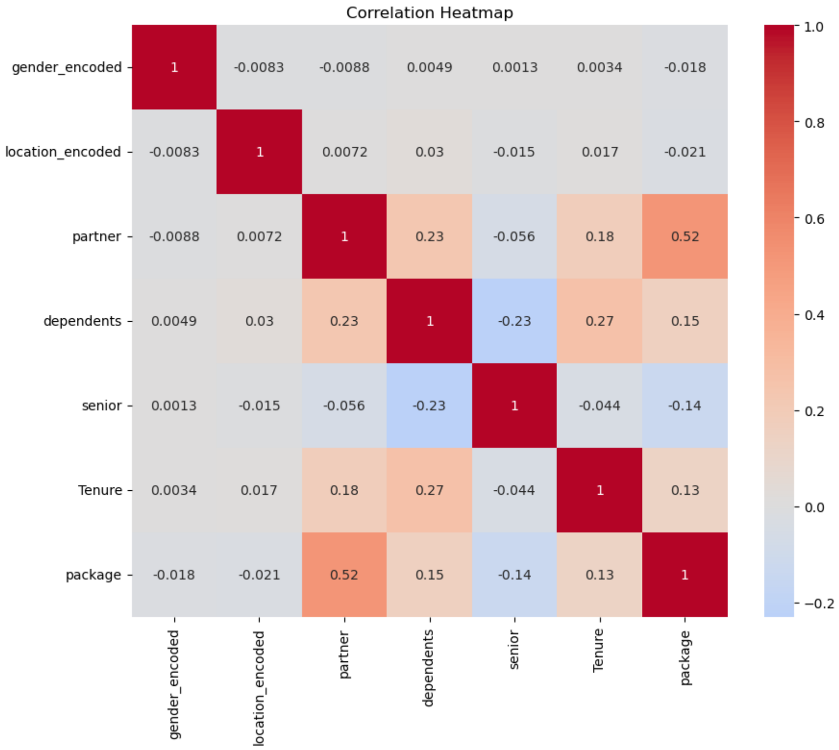

Figure 1: Correlation Matrix heat-map.

To address the issue of improper ’Tenure’ values, a straightforward solution was implemented by replacing them with the median value of 8. This correction was undertaken as ’Tenure’ values should naturally fall within the integer range of 0 to 25. Regarding the ’No reply’ entries in the ’survey’ feature, a pragmatic approach was taken, and these entries were substituted with the calculated mean value of 4.696087084. This replacement resulted in a dataset with a total of 11 unique survey score values, which aligned well with the interpretation of the replaced value as it falls between the ratings of 5 and 4, signifying a moderate level of satisfaction. Finally, Following the necessary data preprocessing steps and EDA, the dataset has been successfully restored to its original size of 7350 observations. The issues related to ’dependents,’ ’Tenure,’ and ’survey’ have been effectively addressed, resulting in a dataset ready for further analysis and modeling.

To conduct feature selection for predicting churn using the refined dataset, a correlation matrix analysis was performed (figure 2). This analysis aimed to identify features that exhibit the strongest correlations with the ’Churn’ class. By examining the correlations between individual features and the ’Churn’ class, we can prioritize those features that have the most significant influence on predicting customer churn. This process assists in identifying key factors that contribute to customer attrition and guides the subsequent steps in model building and analysis.

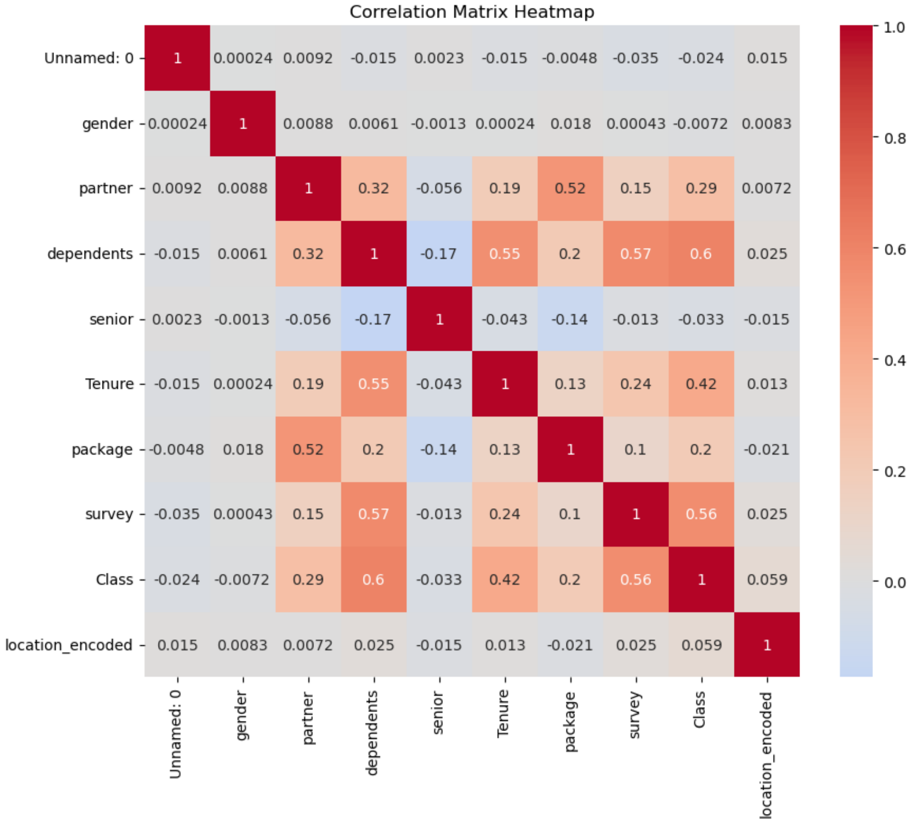

Figure 2: Correlation Matrix heat-map for model feature selection.

Based on the correlation analysis, it has been determined that the features 'dependents', 'survey' and 'Tenure' exhibit the highest correlations with the 'Churn' class. Consequently, these features have been selected for further analysis and modeling, as they are believed to hold significant predictive power in understanding and predicting customer churn behavior. This feature selection process actually helps with modeling process by focusing on the most relevant variables.

3. Developing and testing machine learning models

Various machine learning models were developed, trained, and evaluated for predicting customer churn based on the selected features. The models and their evaluation results are as follows:
1. Decision Tree:
- Accuracy: 0.87
- Precision, recall, and F1-score exhibited good performance for both classes, with slightly higher performance for class 1 (churned customers).

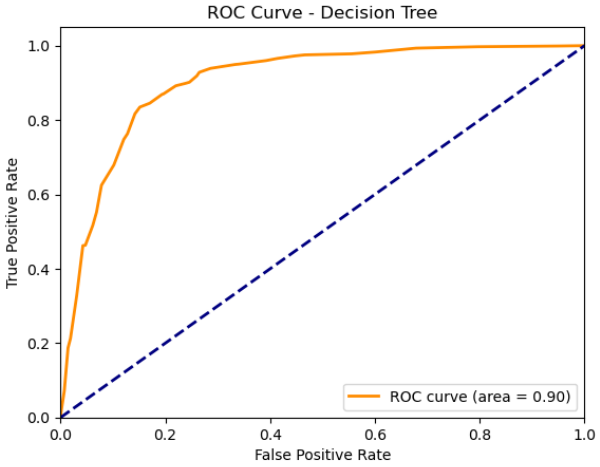

Figure 3: ROC Curve of Decision Tree

2. Random Forest:
- Accuracy: 0.87
- Similar to the Decision Tree, Random Forest achieved balanced precision, recall, and F1-score for both classes.

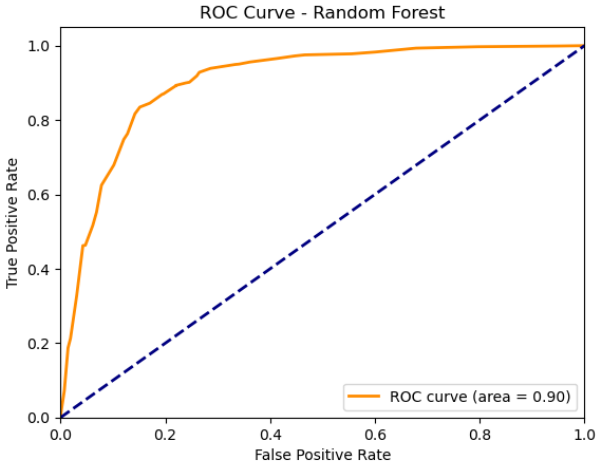

Figure 4: ROC Curve of Random Forest.

3. Logistic Regression:
- Accuracy: 0.87
- Precision and recall showed favorable performance, with a slightly better recall for class 1 (churned customers).

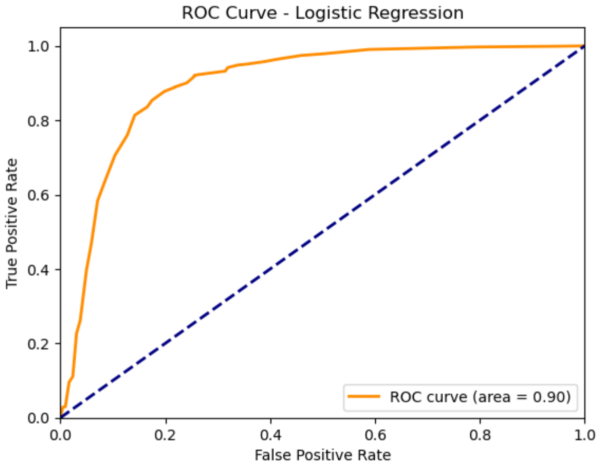

Figure 5: ROC Curve of Logistic Regression

4. Support Vector Machine (SVM):

- Accuracy: 0.87

- Precision, recall, and F1-score were balanced for both classes, with a higher recall for class 1.

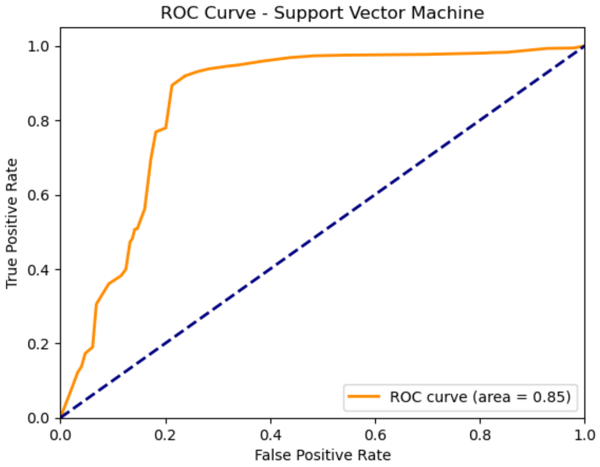

Figure 6: ROC Curve of SVM.

5. k-Nearest Neighbors (k-NN):

- Accuracy: 0.85

- Precision and recall were relatively balanced for both classes, with a slightly better recall for class 1.

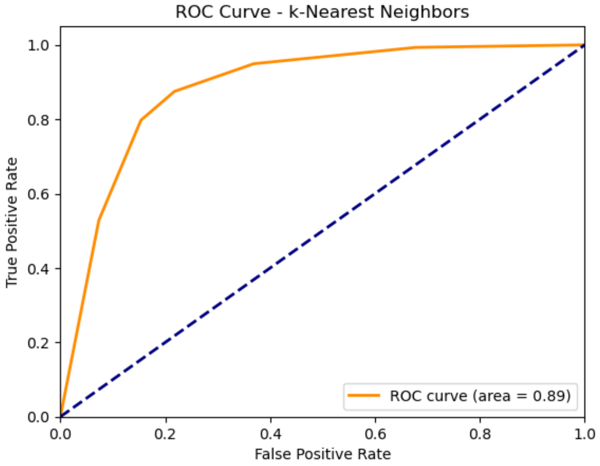

Figure 7: ROC Curve of K-Nearest Neighbors.

6. Deep Neural Network (PyTorch):

- Accuracy: 0.82
- While precision and recall were well-maintained for class 1, precision for class 0 (non-churned customers) was lower, impacting the overall F1-score.

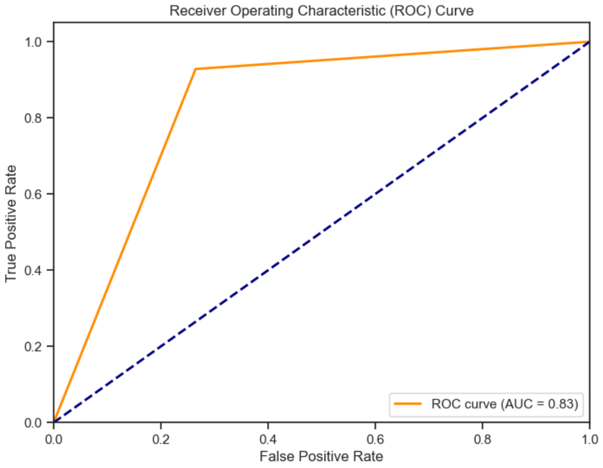

Figure 8: ROC Curve of Deep Neural Network (PyTorch) .

In summary, the ensemble models (Decision Tree, Random Forest) and traditional models (Logistic Regression, SVM, k-NN) exhibited relatively consistent and competitive performance, with an accuracy of around 0.87. The deep neural network, while achieving satisfactory accuracy, displayed some discrepancies in precision for class 0. The choice of model should consider the trade-off between interpretability, ease of implementation, and performance for the specific application.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.87	0.81 / 0.90	0.74 / 0.93	0.77 / 0.91
Random Forest	0.87	0.81 / 0.90	0.74 / 0.93	0.77 / 0.91
Logistic Regression	0.87	0.79 / 0.90	0.74 / 0.92	0.76 / 0.91
Support Vector Machine	0.87	0.79 / 0.91	0.76 / 0.92	0.78 / 0.91
k-Nearest Neighbors	0.85	0.72 / 0.91	0.78 / 0.87	0.75 / 0.89
Deep Neural Network (PyTorch)	0.82	0.91 / 0.81	0.43 / 0.98	0.58 / 0.89

Table 1: Model evaluation results for customer churn prediction. Precision and recall values are shown for both classes (0 and 1) of the Churn variable.

4. Conclusion

The dataset, containing 7350 observations, was described, and the significance of data preprocessing, exploratory data analysis, and subsequent modeling stages was emphasized.

The data preparation phase was instrumental in addressing data anomalies, such as missing values and outliers. Specific attention was given to problematic features like 'dependents,' 'Tenure,' and 'survey,' resulting in a refined dataset of 4608 observations, ready for analysis.

Exploratory data analysis provided valuable insights into data distributions and relationships, aiding in making informed preprocessing decisions. A correlation matrix analysis identified key features for predicting 'dependents,' which led to employing classification models for imputation.

The process of developing and testing machine learning models was meticulously outlined. A range of models, including Decision Tree, Random Forest, Logistic Regression, SVM, k-NN, and a Deep Neural Network, were trained and evaluated. Their performance metrics, such as accuracy, precision, recall, and F1-score, were presented and compared.

In conclusion, the ensemble and traditional models exhibited consistent and competitive performance in predicting customer churn, while the deep neural network displayed some precision discrepancies. The choice of model should consider factors like interpretability and performance, tailored to Lu's Communications' unique needs. The insights gained from this comprehensive analysis equip Lu's Communications with a solid foundation to make informed decisions, minimize churn, and optimize their business strategies. In the context of predicting customer churn, the findings unveiled the Decision Tree model as the most suitable classifier, exhibiting somehow better performance across various metrics. The table below shows the costumers who are more likely to churn, highlighting the churn customers and the cost of keeping them:

Table 2: Churn=yes Customers and their Costs

Customer ID	Cost ($)
I4389	$ 21,915
B9328	$ 41,726
B2668	$ 35,151
G4357	$ 25,340
Total	$ 124,132