1.0 Project Introduction

The Telco Customer Churn dataset provides customer information regarding a fictitious telecommunications company operating in California. This dataset encompasses records for 7,043 customers, delineating customers who have discontinued, retained, or newly subscribed to their services. The data was originally released by IBM, and the version used for this exercise can be accessed on Kaggle - https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

The overarching goal of our project is to predict customer churn in the dataset using three machine learning models—Logistic Regression, Random Forest, and an Artificial Neural Network (ANN). Each model has been applied to analyse the same dataset, with the objective of identifying which customers are likely to leave (churn) the service. The performance of these models was evaluated using a variety of metrics such as accuracy, precision, recall, f1-score, and the area under the The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). Additionally, an analysis of feature importance was conducted to gain a deeper understanding of the factors contributing to customer churn. We start by importing our dataset into a Pandas DataFrame within our Jupyter Notebook for exploratory data analysis (EDA). A separate Jupyter Notebook is provided for the EDA, as well as for each machine learning (ML) method applied in this project. After loading the dataset, we use the '.head()' function to examine the first few rows of the dataset. This step helps us initially understand the data's structure, including the total rows and columns and the variable types (numerical, categorical, etc.). The dataset's columns feature diverse customer data, such as demographic details, account charges, and a churn indicator to show if the customer has churned.

1.1 Understanding & Cleansing the data

Initially, the dataset summary reveals that there are no missing values in any of the columns. However, it's worth noting that the 'TotalCharges' column is of type 'object' (string), suggesting the presence of non-numeric values or blanks that were not recognised as missing values. Additionally, the 'SeniorCitizen' column serves as a binary indicator (0 or 1), with a mean value close to 0.16, indicating that roughly 16% of customers using the service are classified as senior citizens. After converting the 'TotalCharges' column from an object type to a numeric type, we discovered that 11 instances resulted in missing values. Subsequently, we removed the rows with missing values in the 'TotalCharges' column leaving us with a cleansed dataset of 7032 entries or customers. With our dataset now devoid of these missing values, we can conduct more accurate analysis and modelling.

1.2 Statistical Analysis & Data Visualisation

Statistical analysis and visualisation offer further insights into our dataset. Customer tenure ranges from 0 to 72 months, with the average tenure calculated to be approximately 32 months. The wide range suggests a diverse customer base, ranging from new subscribers to long-standing customers, which might reflect varying levels of loyalty. Examining the 'MonthlyCharges' column reveals charges ranging from $18.25 to $118.75, indicating a broad spectrum of service plans among customers. Calculating the mean monthly service charge from the dataset yields $64.80. The 'TotalCharges' column shows charges ranging from $18.80 to $8684.80 with a mean total charge of $2283.30. The standard deviation for the total charges is $2266.77 indicating that there is considerable variability in the total charges across the dataset.

We analyse the distribution of categorical variables, specifically focusing on the 'Churn' variable, to identify initial patterns. This examination helps us understand how these variables relate to customer churn. Upon examining the distribution of customer churn, we observed that approximately 26.6% of customers have churned, while roughly 73.4% have not. This suggests a notable churn rate that the company may wish to address. Subsequently, through bar charts, we further investigate how key categorical variables in the dataset impact churn rates. The distribution of churn seems to be relatively balanced across genders, suggesting that gender may not be a significant predictor of churn for the dataset. However, we observed that senior citizens have a higher churn rate compared to non-senior citizens, indicating that age could indeed play a role in churn likelihood.

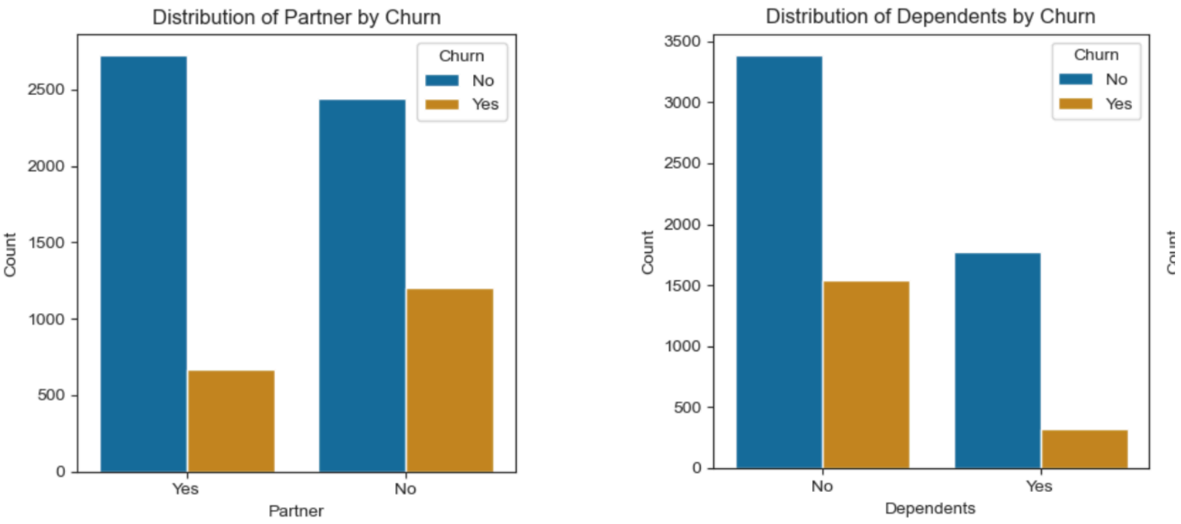

Figure 1:(a) Distribution of Churn by Partner Status: A bar chart displaying the churn distribution among customers with and without partners.(b) Distribution of Churn by Dependents: A bar chart illustrating how churn varies among customers with and without dependents

Furthermore, as shown in Figure 1(a), our analysis reveals that customers without partners tend to churn more than those with partners. This insight is valuable as it suggests that customers with partner commitments may perceive more value in the company’s service, potentially leading to higher retention rates among this demographic. Similarly to the 'Partner' variable, as shown in Figure 1(b), customers without dependents churn at a higher rate than those with dependents. Additionally, our analysis revealed a higher churn rate among customers with fibre optic internet service compared to users with DSL or no internet service. This discrepancy could be attributed to factors such as pricing or service reliability. Furthermore, we observed that customers on month-to-month contracts exhibit a significantly higher churn rate compared to those on one-year or two-year contracts. This observation underscores the causal relationship between contract length and customer retention, suggesting that longer contract durations contribute to higher customer loyalty.

1.3 Feature Engineering

We conducted outlier analysis on our dataset employing the Interquartile Range (IQR) method, which identifies outliers as data points that lie more than 1.5 times the IQR distance from the quartiles. Our analysis specifically targeted the 'tenure', 'MonthlyCharges', and 'TotalCharges' columns, while generally excluding columns with binary or categorical variables from outlier detection. When outliers are identified, our aim is to remove the rows where any outliers are present. This process aims to ensure that our dataset remains consistent and avoids having incomplete data for any given observation.

After applying the IQR method to identify and remove outliers from the continuous numerical columns, we find that the shape of our dataset is unchanged (7032 rows and 21 columns). This outcome suggests that the data within these columns may not contain extreme values that fall outside the typical 1.5 times IQR range from the quartiles. Box plots provided a further insight into the spread of the distribution and consolidate our outlier detection method through IQR. Having gained a comprehensive understanding of our dataset through EDA, the next phase of our project involves developing a tool which leverages Machine Learning and Deep Learning techniques to predict churn outcomes.

2.0 Logistic Regression

For our project, the first machine learning method we explore is logistic regression. This analysis models the probability of a discrete outcome, such as churn, based on input variables.[1] The logistic regression model can be a popular method for binary classification tasks, like predicting customer churn as it works by estimating probabilities using a logistic function, which is particularly useful for binary outcomes.

2.1 Data Preprocessing & Encoding

The initial step in our analysis involves encoding the dataset's variables. We transform these variables into 'dummy' or 'indicator' variables using the 'pd.get_dummies()' function. This process is necessary as logistic regression, like many other machine learning algorithms, necessitates a numerical input. We omit the 'Churn' and 'customerID' columns from the encoding process. 'Churn' serves as our target variable, while 'customerID', being a unique identifier, offers no predictive value. We convert the 'Churn' column into a binary format where customers who have churned are marked as 1 ('Yes') and those who have not as 0 ('No'). This conversion (Target Variable Transformation) is crucial for logistic regression, which is designed to predict binary outcomes.

2.2 Data Splitting & Model Training

We split the dataset into training and testing sets using a 70-30 ratio, where 70% of the data is used for training the model, and 30% is reserved for testing the model performance. We use the same ratio for our Random Forest (see Section 3) and Artificial Neural Networks (see Section 4) models to accurately compare performance on our broader unseen test dataset across the three models. The ‘random_state parameter’ is initialised to ensure that our results are reproducible, which means the same split occurs every time the code is run. We keep the ‘random state’ value constant across all three models to also compare performance and output features. We scale the features for our model using the ‘StandardScaler’ feature to ensure that all features contribute equally to the model. This is particularly beneficial for features with different measurement units or significant scale variations, as these factors can affect the model's convergence and performance. We initialise and train our logistic regression model with a high number of maximum iterations (1000) to ensure convergence.

2.3 Model Prediction & Testing

Our Logistic regression model demonstrates good predictive ability, particularly for identifying customers who have not churned, although there seems to be room for improvement in identifying churned customers more accurately. The model's accuracy is approximately 73.9%, indicating that it correctly predicts whether a customer has churned or not about 74% of the time across the test dataset. Precision measures the proportion of true positive predictions in all positive predictions, while recall measures the proportion of true positive predictions in all actual positives. For customers who have not churned (label 0), the precision is 0.74, and the recall is 0.91, which usually indicates high performance. For customers who have churned (label 1), precision is 0.51, and recall is 0.80, which is lower but reasonable. The F1-score, which acts as the harmonic mean of precision and recall, provides a single metric to assess the balance between them. Our logistic regression model achieves an F1-score of 0.80 for non-churned customers and 0.62 for churned customers, reflecting its overall effectiveness but also potentially highlights areas for improvement, particularly in correctly identifying churned customers.

The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) is a comprehensive metric we also used to evaluate the performance of our model. The AUC measures the model's ability to discriminate between positive (churn) and negative (no churn) classes. The AUC for our Logistic Regression analysis is calculated as 0.84. An AUC value ranges from 0 to 1, where 0.5 indicates no discriminative ability (equivalent to random guessing). A value closer to 1 indicates a high ability to correctly discriminate between positive and negative classes. Values closer to 0 indicate that the model is performing worse than random guessing, often meaning it is inversely predicting the outcomes. We also examined and visualised the feature importances derived from our logistic regression model. The analysis reveals that 'tenure' is possibly the most critical feature, indicating that the duration a customer has been with the company significantly influences the model's predictions. In customer churn analysis, this suggests that the length of service is a key factor in predicting the likelihood of churn.

3.0 Random Forest

The second machine learning approach we examine is the Random Forest algorithm.[2] By aggregating predictions from multiple decision trees, this method synthesises a unified outcome. Through the collective insights of multiple trees, the Random Forest technique seeks to enhance prediction accuracy and robustness. It achieves this by overcoming the limitations of individual decision trees through a strategy known as ensemble learning.

3.1 Data Preprocessing & Encoding

Mirroring the approach taken with our Logistic Regression model, we prepare the feature set ‘X’ for our Random Forest model by excluding the target variable 'Churn' and the non-predictive 'customerID' identifier. The variables are subsequently transformed into 'dummy' variables, rendering the data appropriate for Random Forest analysis. Additionally, the target variable 'y' is extracted from the 'Churn' column by converting 'Yes' to 1 and all other responses to 0, facilitating binary classification. Although Random Forest algorithms, owing to their decision tree structure, handle features of varying scales, we apply scaling to ensure consistency. This approach is particularly relevant as we are concurrently experimenting with another algorithm that necessitates scaled features. (Refer to Section 2)

3.2 Data Splitting & Model Training

A Random Forest Classifier, comprising 500 trees (n_estimators=500), is initialised and subsequently trained using the scaled training dataset. Opting for a relatively high number of trees is intended to enhance the model's accuracy and robustness, albeit potentially increasing computational complexity. After training, this model is used to make predictions on the scaled test dataset. Consistent with our methodology across all three models, we maintain a 70-30 split for dividing the dataset into training and testing sets, using the same random state value (42) to ensure comparability.

3.3 Model Prediction & Testing

The model attains an accuracy of approximately 78%. In identifying customers who do not churn (class 0), the model exhibits high precision (0.82) and recall (0.90), indicating strong accuracy in predicting true negatives. However, for predicting churn (class 1), precision decreases to 0.62, and recall to 0.46, revealing some difficulties in precisely identifying churn cases. The F1-scores, reflecting a balance between precision and recall, stand at 0.86 for non-churn predictions and 0.53 for churn predictions. This underscores the model's greater effectiveness in accurately identifying non-churn cases compared to churn cases. The Random Forest model demonstrates strong performance in predicting non-churn customers; however, it faces challenges in accurately identifying churn customers, as indicated by lower precision and recall scores for the churn class. An AUC of 0.81 indicates that the Random Forest model possesses a strong degree of separability, enabling it to effectively distinguish between the two classes (churn and no churn).

The identification of "Total Charges" as the most important feature in a Random Forest analysis for predicting customer churn provides valuable insights into customer behaviour and informs retention strategies. In the context of this analysis, where the Random Forest model achieved an Area Under the ROC Curve (AUC) of 0.81, indicating good model performance, the prominence of "Total Charges" may suggest key interpretations and implications. The total amount spent could also reflect the level of service usage. Customers who use a wide range of services or subscribe to higher-tier plans are likely contributing more to the total charges and might have a different relationship with the service provider. This variance in usage and investment could potentially influence their decision regarding whether to churn. Hence, understanding the significant role of total charges in predicting churn could empower the company to tailor their retention strategies more effectively.

4.0 Artificial Neural Network

Artificial Neural Networks (ANNs) are computational models that recognise patterns and make decisions by emulating the signalling between neurons in animal brains.[3] Therefore, in the context of predicting telecom customer churn, ANN could have the ability to identify complex patterns and relationships within customer data that traditional statistical methods may not easily detect.

4.1 Data Preprocessing & Encoding

ANNs are adept at learning to recognise patterns in various aspects of customer data, including behaviour usage, satisfaction levels, and other indicators of potential churn. Given that customer data is often high-dimensional (comprising many features) and nonlinear, ANNs excel at extracting valuable insights from these complex datasets, minimising the need for extensive feature engineering. Similar to the data preprocessing steps in the two previously discussed machine learning methods, the dataset is transformed such that categorical variables are converted into dummy variables for numerical analysis. The target variable, 'Churn', is encoded as a binary variable (y), with 'Yes' corresponding to 1 (indicating churn) and 'No' to 0 (indicating no churn). Additionally, the 'customerID' column is also removed due to its lack of contribution to the model's predictive capability.

4.2 Neural Network Model Setup & Training

We divide the dataset into training and test sets, with 70% allocated for training and 30% for testing. Moreover, the features are scaled using the StandardScaler to ensure uniformity in data magnitude. Using TensorFlow and its high-level API, Keras, we construct a Sequential model comprising three Dense layers. This architecture is specifically tailored for binary classification tasks.

The final layer employs a sigmoid activation function to produce probabilities, indicating the likelihood of churn. The model is compiled using the Stochastic Gradient Descent (SGD) optimiser and is trained with binary cross-entropy as the loss function, aptly reflecting the binary nature of the classification task. The training process includes the adjustment of the model's weights based on the computed class weights, ensuring balanced learning despite class imbalance.

4.3 Model Prediction & Testing

An accuracy of 74% demonstrates that the model accurately predicts the churn status in approximately three-quarters of cases, reflecting a satisfactory performance level for the intricate task of customer churn prediction. The model exhibits higher precision in predicting no-churn cases (0.82) compared to churn predictions (0.57), indicating a high reliability in identifying customers who are less likely to churn.

Furthermore, the recall scores indicate that the model is more adept at identifying true no-churn instances over churn instances, reinforcing its strength in distinguishing customers who will continue their service. The f1-score indicates that the Artificial Neural Network (ANN) performs better in predicting cases of no churn. The AUC of 0.75 for our ANN signifies a good but not excellent capability in distinguishing between customers who will churn and those who will not. By incorporating TensorFlow and deep learning techniques, we enable the creation of advanced model architectures capable of detecting complex patterns in customer behaviour, significantly enhancing our predictive capabilities. Nevertheless, the results highlight areas for enhancement, particularly in improving churn prediction accuracy. To refine the model could involve further tuning of the neural network architecture, exploring more advanced feature engineering, or employing techniques like oversampling the minority class.

5.0 Performance Metrics Analysis & Discussion

It's important to highlight that, in our study, tackling the challenge of class imbalance was essential across all the three predictive models to guarantee robust and equitable evaluation and performance during the model setup and training phases. Class imbalance occurs in situations where the instances one class significantly outnumbers those in the opposite class, which can cause the model to favour the majority class, leading to biased predictions. To address this issue, we implemented a variety of techniques specifically chosen to complement the unique characteristics and mechanics of each model as detailed in our Jupyter Notebooks.[4][5] Upon comparison of our models, as detailed in Table 1, the Random Forest algorithm exhibits the highest overall accuracy. This indicates its superior capability in accurately classifying customers into churned and not churned categories across the entire dataset. Furthermore, the Random Forest model demonstrates the highest precision in predicting churn among the evaluated models. This implies a lower incidence of false positives , where customers are incorrectly predicted to churn when they, in fact, do not, compared to the other two models. Logistic Regression exhibits the highest recall for churn prediction, marking it is as the most effective model in correctly identifying actual churn cases, albeit with a potential increase in false positives, as indicated by its lower precision. Additionally, it achieves the highest F1-score for churn, reflecting a superior balance between precision and recall relative to the other models. This efficiency suggests that Logistic Regression is adept at pinpointing churn cases while minimising the misclassification of non-churn cases as churn.

Table 1: Comparative Analysis of Key Performance Metrics Across Three Predictive Models for Customer Churn

Random Forest stands out in terms of overall accuracy and precision, rendering it an excellent choice when the primary objective is to minimise false positives. Logistic Regression distinguishes itself with its high recall and F1-score for churn prediction, indicating its effectiveness in capturing as many true churn cases as possible, albeit with an increased incidence of false positives. Although the Artificial Neural Network provides a balanced approach, it does not excel in any specific metric when compared to the other three models. Logistic Regression exhibits the highest AUC value, scoring at 0.84, which signifies its superior ability to differentiate between customers likely to churn and those likely to remain. This indicates that the Logistic Regression model achieves an effective balance between sensitivity (true positive rate) and specificity (true negative rate), standing out among the three models. With an AUC of 0.83, Random Forest also demonstrates a robust discriminative capability, closely rivalling Logistic Regression and indicating its strong ability to distinguish between customers who will churn and those who will not. This consistent performance aligns with the model's previously noted high accuracy and precision in churn prediction. The Artificial Neural Network, with an AUC of 0.76, demonstrates a reasonable capacity to differentiate between the two classes. However, this indicates that it is the least effective among the three models in this particular aspect. Despite this, an AUC of 0.76 exceeds the performance of random chance, indicating that with additional tuning or adjustments to the network architecture, there's potential to enhance its performance significantly.

6.0 Conclusion & Improvements

In summary, while all three models are capable of distinguishing between churned and non-churned customers, Logistic Regression and Random Forest outperform the Artificial Neural Network across most metrics, particularly in terms of the AUC metric. Logistic Regression, in particular, exhibits the most robust performance for the specific task of churn prediction, adeptly balancing the need to accurately identify churn cases with the overall capability to classify customers. When choosing a model for practical deployment, it's crucial to consider not only the AUC but also additional performance metrics and the business context. This evaluation should include considering the costs associated with false positives, such as unnecessary interventions for customers incorrectly predicted to churn, versus the costs of false negatives, which involve missing opportunities to intervene with customers at risk of churning. The decision can also be swayed by the model's interpretability, scalability, and how seamlessly it integrates into existing systems.To enhance our analysis, particularly in the context of machine learning and deep learning projects such as customer churn prediction, several strategies could further refine our approach. Initially, expanding our feature sets should be a priority. Investigating additional data sources could introduce new, relevant features for predicting churn, such as customer feedback scores or interactions with customer support. Furthermore, beyond the models already employed, exploring other algorithms that may better capture the nuances of our dataset should be considered. Algorithms like Gradient Boosting Machines (GBM), Extreme Gradient Boosting (XGBoost), or LightGBM may offer advanced capabilities that could potentially improve our predictive accuracy.[6] Additionally, employing optimisation techniques such as Principal component analysis (PCA), grid search, random search, Bayesian optimisation, or genetic algorithms to identify the optimal model parameters could markedly enhance model performance.[7] Implementing more robust cross-validation techniques, such as stratified k-fold or time-series cross-validation, could also be beneficial.[7][8] In light of our focus on predicting customer churn using machine learning and deep learning, advancing our models through experimentation with various deep learning architectures is pivotal. The integration of dropout layers can help prevent overfitting, while optimising activation functions can further refine the learning process. Beyond model enhancement, the practical application of these predictions in real-world

Model	Churn (0 = ‘No’,1 = ‘Yes’)	Precision	Recall	F1-score	AUC	Accuracy
Logistic Regression	0	0.74	0.91	0.80	0.84	0.74
	1	0.51	0.80	0.62		
Random Forest	0	0.82	0.90	0.86	0.81	0.78
	1	0.62	0.46	0.53		
Artificial Neural Network	0	0.86	0.78	0.82	0.75	0.74
	1	0.51	0.65	0.57		

scenarios is crucial. Launching a pilot program with actual customers to test the efficacy of proposed retention strategies, guided by our predictive models, offers invaluable insights. This hands-on approach not only validates our model's predictions but also allows for iterative refinement of strategies based on real customer responses.

7.0 References

[1] Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression* (Vol. 398). John Wiley & Sons.
[2] Breiman, L., 2001. Random forests. *Machine learning*, *45*, pp.5-32.
[3] Zou, J., Han, Y. and So, S.S., 2009. Overview of artificial neural networks. *Artificial neural networks: methods and applications*, pp.14-22.
[4] Kotsiantis, S., Kanellopoulos, D. and Pintelas, P., 2006. Handling imbalanced datasets: A review. *GESTS international transactions on computer science and engineering*, *30*(1), pp.25-36.
[5] Spelmen, V.S. and Porkodi, R., 2018, March. A review on handling imbalanced data. In 2018 international conference on current trends towards converging technologies (ICCTCT) (pp. 1-11). IEEE.
[6] Sibindi, R., Mwangi, R.W. and Waititu, A.G., 2023. A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. *Engineering Reports*, *5*(4), p.e12599.
[7] Fushiki, T., 2011. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*, pp.137-146.
[8] Bergmeir, C. and Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, pp.192-213.