Customer Churn Modelling

Geo M Benny

Introduction

In today's competitive telecom industry, customer retention has become one of the most critical business challenges. Customer churn—the phenomenon where customers discontinue using a company's services—is a significant concern for telecom companies globally. The telecom sector, characterised by intense competition, rapid technological advancements, and a large customer base, sees millions of subscribers switching service providers due to various reasons such as dissatisfaction with services, better offers from competitors, or changes in personal preferences. As acquiring new customers is often more expensive than retaining existing ones, reducing churn has a direct impact on profitability and operational efficiency.

This project focuses on building a predictive model to identify customers at risk of churning. Predicting churn enables companies to take proactive steps to retain customers by offering personalised services, better plans, or resolving issues that lead to dissatisfaction. By understanding and forecasting customer churn behaviour, companies can devise data-driven strategies to enhance customer satisfaction, improve loyalty, and ultimately, retain customers more effectively.

Problem Statement

The problem addressed in this project revolves around predicting whether a customer will churn or not, based on their usage patterns and interactions with the telecom service provider. A high churn rate not only leads to revenue loss but also affects the company's market share. The aim is to utilise machine learning techniques to predict which customers are more likely to discontinue their service, based on historical data such as call durations, international and voicemail plan subscriptions, and customer service interactions.

To build a reliable churn prediction model, various machine learning algorithms were considered, including K-Nearest Neighbours (KNN) and Artificial Neural Networks (ANN). These models were trained using a customer dataset that contains various attributes reflecting customer behaviours and service usage. The goal was to develop models that can accurately predict whether a customer is at risk of churning, thus allowing the telecom company to intervene early and reduce churn rates.

Objectives

The primary objective of this project is to develop predictive models that can accurately classify customers as "churners" or "non-churners" based on their historical data. Specific objectives include:

- **Data Analysis**: Perform exploratory data analysis (EDA) to understand key trends and patterns that correlate with customer churn. This involves investigating the distribution of key variables such as call durations, customer service interactions, and plan subscriptions.
- **Data Preprocessing**: Apply appropriate preprocessing steps to handle missing values, encode categorical variables, and scale features, ensuring that the data is in a suitable format for training machine learning models.
- **Model Development**: Implement machine learning models such as KNN and ANN to predict churn, tuning model hyper parameters to maximise predictive accuracy. The rationale behind choosing these models stems from their ability to handle complex classification tasks, where identifying patterns among customer behaviours is essential.
- **Model Evaluation**: Evaluate model performance using metrics such as accuracy, precision, recall, and F1 score. This ensures that the models are not only accurate but also robust in predicting true positives (churners) and minimising false positives (non-churners incorrectly classified as churners).
- **Deployment**: Convert the trained models into a deployable format (.pkl) to be showcased on a web application using Streamlit, making the predictions accessible to end-users in a user-friendly interface.

Overview of Dataset

The dataset used in this project comprises customer data from a telecom company, with records containing information about customer demographics, service usage, and interactions with the company. The dataset consists of **5000** records and **18** features, with each feature providing valuable insights into customer behaviour and service usage patterns. Key variables include:

- **State**: Geographical information about the customer's location.
- **Account Length**: Duration of time the customer has been with the company.
- **Area Code**: The area code associated with the customer's phone number.
- **International Plan and Voicemail Plan**: Whether the customer has opted for international calling and voicemail services.
- **Service Usage**: This includes variables reflecting the total duration and number of calls made during the day, evening, night, and on international calls. Corresponding charges for these services are also included.
- **Customer Service Interactions**: The number of times a customer has contacted the company's customer service department, which could indicate dissatisfaction with the service.

The project utilises a supervised learning approach where the target variable is binary—whether a customer has churned or not (1 for churners, 0 for non-churners).

account_length	area_code	international_plan	voice_mail_plan	number_vmail_messages	total_day_minutes	total_day_calls	total_day_charge	total_eve_minutes	total_eve_calls	total_eve_charge	total_night_minutes	total_night_calls	total_night_charge	total_intl_minutes	total_intl_calls	total_intl_charge	number_customer_service_calls	churn
107	area_code_415	no	yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.7	1	no
137	area_code_415	no	no	0	243.4	114	41.38	121.2	110	10.3	162.6	104	7.32	12.2	5	3.29	0	no
84	area_code_408	yes	no	0	299.4	71	50.9	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	no
75	area_code_415	yes	no	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	no
121	area_code_510	no	yes	24	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	2.03	3	no
147	area_code_415	yes	no	0	157.0	79	26.69	103.1	94	8.76	211.8	96	9.53	7.1	6	1.92	0	no
117	area_code_408	no	no	0	184.5	97	31.37	351.6	80	29.89	215.8	90	9.71	8.7	4	2.35	1	no
141	area_code_415	yes	yes	37	258.6	84	43.96	222.0	111	18.87	326.4	97	14.69	11.2	5	3.02	0	no
65	area_code_415	no	no	0	129.1	137	21.95	228.5	83	19.42	208.8	111	9.4	12.7	6	3.43	4	yes
74	area_code_415	no	no	0	187.7	127	31.91	163.4	148	13.89	196.0	94	8.82	9.1	5	2.46	0	no
168	area_code_408	no	no	0	128.8	96	21.9	104.9	71	8.92	141.1	128	6.35	11.2	2	3.02	1	no

The above image is the preview of the dataset.

Data Overview

The dataset used in this project is central to understanding customer behaviour and making accurate predictions about customer churn in the telecom sector. The data is composed of various customer attributes that reflect their service usage patterns, account information, and interaction history with the company. By analysing these attributes, we can extract meaningful insights into which factors contribute most to churn and build robust predictive models accordingly.

Data Composition

The dataset consists of 5000 **records** and **20 features**, with each record corresponding to a unique customer. The features include both numerical and categorical variables that describe different aspects of customer behaviour and service usage. Here's a breakdown of the key features in the dataset:

- **id**: A unique identifier assigned to each customer. This field is primarily used for tracking individual customers but does not contribute to model training since it holds no predictive value.
- **state**: A categorical variable representing the state or region where the customer resides. Geographic information such as this can sometimes reveal location-based patterns in customer behaviour and churn risk.
- **account_length**: A numerical variable that measures the length of time (in days) the customer has been with the company. This variable is essential as long-term customers may exhibit different churn patterns compared to newer customers. For instance, newer customers might be more susceptible to churn if their expectations are not met early on.
- **area_code**: A categorical variable representing the area code associated with the customer's phone number. While this variable might seem redundant at first glance, it can help detect regional differences in customer behavior.
- **international_plan**: A binary categorical variable indicating whether the customer has subscribed to an international calling plan (Yes/No). Customers with international plans may have unique usage patterns and could be less likely to churn if they frequently make international calls.
- **voice_mail_plan**: Another binary categorical variable that indicates whether the customer has opted for a voicemail plan (Yes/No). Voicemail plans might be a reflection of customer preferences for enhanced services and can influence customer satisfaction and retention.
- **number_vmail_messages**: A numerical variable indicating the total number of voicemail messages the customer has received. This feature is closely

linked to the voice_mail_plan variable and provides additional insights into how often the customer utilises their voicemail services.

- **total_day_minutes**, **total_day_calls**, **total_day_charge**: These three numerical variables describe the customer's total call duration, the number of calls made, and the corresponding charges during daytime hours. Daytime usage can vary widely among customers and is an essential indicator of their engagement with the service.
- **total_eve_minutes**, **total_eve_calls**, **total_eve_charge**: Similar to the day-related features, these variables represent the customer's total call duration, number of calls, and charges during evening hours. The comparison between daytime and evening usage can reveal different customer usage patterns that could be related to churn.
- **total_night_minutes**, **total_night_calls**, **total_night_charge**: These variables measure call duration, the number of calls, and the corresponding charges during nighttime hours. Some customers may exhibit heavier usage during specific times of the day, and this could impact their likelihood of churning based on satisfaction with service quality during these periods.
- **total_intl_minutes**, **total_intl_calls**, **total_intl_charge**: These three numerical features capture the customer's international call duration, the number of international calls made, and the associated charges. International calling behaviour is a key factor for customers with international plans and can significantly influence customer retention.
- **number_customer_service_calls**: A numerical variable indicating the total number of times the customer has contacted the company's customer service department. This feature is often a strong predictor of churn, as frequent customer service interactions may indicate dissatisfaction with the service or unresolved issues.

Target Variable

The target variable in this dataset is **churn**, a binary indicator of whether or not the customer has discontinued their service with the telecom company. A value of 1 indicates that the customer has churned, while a value of 0 indicates that the customer has remained with the company. Predicting this target variable is the primary objective of the machine learning models developed in this project.

Summary Statistics

Initial exploration of the dataset revealed several important summary statistics for both numerical and categorical features:

- **Account Length**: The average account length for customers in the dataset is approximately **100-150 days**, with a range that varies from customers who have just joined to those who have been with the company for several years. This variation in account length provides a diverse range of customer experiences to analyse.

- **Total Day Minutes**: On average, customers spend **200-300 minutes** on daytime calls each month. The total daytime call charges are closely correlated with call duration, given the telecom company's billing structure.
- **Customer Service Calls**: The majority of customers have made fewer than **5 calls** to customer service, with a smaller proportion having made more frequent calls. High levels of customer service interaction can often signal frustration or dissatisfaction, which might lead to a higher likelihood of churn.

For categorical variables like state, international_plan, and voice_mail_plan, frequencies were calculated to understand the distribution of customers across these categories. For example, a large proportion of customers are concentrated in specific states, while fewer customers have opted for international or voicemail plans. These distributions help inform the modelling process, as certain categories may be more predictive of churn than others.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical step in understanding the patterns and relationships within a dataset before building predictive models. During the EDA for this telecom customer churn dataset, several important insights were uncovered that helped guide the development of machine learning models and feature engineering. Through visualisations such as histograms and heatmaps, and statistical analysis, we were able to identify key trends that influenced customer churn.

Call Duration Patterns

One of the primary features analysed was the call duration across different periods of the day, namely the **total_day_minutes**, **total_eve_minutes**, and **total_night_minutes** variables. It was observed that customer usage was highest during nighttime hours, as reflected by the **total_night_minutes** variable, which had the largest range and average values compared to day and evening periods. The average night time call duration ranged between **200 to 300 minutes** for most customers, while evening and day calls were generally shorter in duration.

This distribution indicates that customers rely more on telecom services during the night, which may suggest that night time usage patterns could play a critical role in customer satisfaction and potentially impact churn. For example, customers who experience poor service during peak night time hours might be more inclined to switch to a different provider. Conversely, day and evening periods appeared to be less crucial in determining customer churn, though their impact was still considered in the overall analysis.

Histograms of call durations for these different periods showed a skewed distribution, with a majority of customers having moderate usage and a few exhibiting significantly higher usage. This insight prompted further analysis to understand if extreme usage in any period was predictive of churn behaviour.

Customer Service Interactions

Another key finding from the EDA was the relationship between customer churn and interactions with customer service, as captured by the **number_customer_service_calls** variable. Initial analysis revealed a clear trend: customers who had a higher number of customer service calls were more likely to churn. Customers with **more than 3 customer service interactions** showed a noticeably higher churn rate than those with fewer interactions.

This trend suggests that frequent contact with customer service may be indicative of dissatisfaction with the service. Whether due to billing issues, service outages, or other unresolved problems, customers who reach out to customer service multiple times are often frustrated, which increases the likelihood that they will seek alternatives. This finding aligns with industry norms, where poor customer service experiences are a significant driver of churn.

To visualise this relationship, a bar plot was generated, showing the proportion of churned customers against the number of customer service interactions. The plot clearly illustrated an upward trend, reinforcing the importance of addressing customer service issues proactively to reduce churn.

International Plan Insights

Another interesting insight came from examining the **international_plan** variable. Customers who subscribed to an international plan exhibited different usage behaviours compared to those who did not have the plan. Specifically, international plan customers tended to have longer call durations for both domestic and international calls, and their overall usage patterns differed significantly.

While customers with international plans generally represented a smaller segment of the customer base, their unique usage patterns had a noticeable impact on churn. Interestingly, customers with international plans showed a lower churn rate, potentially because these customers relied more heavily on the service for international communication and thus were less likely to switch providers. This finding underscores the potential of offering specialised plans to retain high-value customers.

Conclusion of EDA

In summary, the EDA phase revealed several important insights about customer behaviour and churn patterns in the telecom sector. Night time call duration emerged as a key area of usage, with higher activity during this period potentially linked to churn behaviour. Frequent customer service interactions were found to be strongly correlated with churn, highlighting the importance of maintaining high-quality customer service. Additionally, customers with international plans demonstrated unique behaviours and lower churn rates, suggesting that personalised services could be an effective retention strategy.

These findings provided a solid foundation for model development, guiding feature selection, and informing decisions around data preprocessing and engineering. The visualisations, distributions, and correlation analysis not only helped in better understanding the dataset but also played a crucial role in shaping the subsequent machine learning efforts to predict customer churn accurately.

Data Preprocessing

Effective data preprocessing is vital to ensure that the machine learning models receive clean, consistent, and correctly formatted data. The quality of the input data greatly influences the performance of the predictive models, and as such, several preprocessing steps were applied to the dataset used for the customer churn model. These steps included handling missing values, encoding categorical variables, feature scaling, and feature engineering. Each of these processes is discussed in detail below.

Handling Missing Values

The first step in data preprocessing involved addressing any missing values within the dataset. Missing data can pose significant challenges to model accuracy and reliability, as it may lead to skewed insights or misrepresentations of customer behavior. To identify any potential issues, the .isnull().sum() function was used to check for missing values across all variables in both the training and test datasets.

Upon inspection, the dataset did not contain any missing values, which simplified the process. However, had there been any missing data, appropriate strategies such as imputation (using the mean, median, or mode) or deletion (if the data was not critical or was sparse) would have been applied. For example, continuous variables might be imputed using the median value, while categorical variables could be replaced using the mode, ensuring that data integrity is maintained without introducing bias.

In this case, the absence of missing data meant that no imputation or deletion was necessary, allowing for a more straightforward data preparation process.

Encoding Categorical Values

Many machine learning algorithms require numerical input, making it necessary to encode categorical variables. This dataset included several categorical features, such as **state**, **international_plan**, and **voice_mail_plan**, which needed to be converted into numeric representations.

The **state** feature, which included various state abbreviations, could not be used in its raw form since machine learning algorithms do not process text-based inputs. For simplicity, label encoding was applied to the **state** variable, which converted each unique state into a corresponding numeric label. While label encoding is suitable for this purpose, one-hot encoding could also have been considered if the goal were to avoid introducing ordinal relationships between the states.

In contrast, the **international_plan** and **voice_mail_plan** variables contained binary values ('yes' or 'no'), which made them easier to encode. These variables were mapped to numeric values using binary encoding, where 'yes' was mapped to 1 and 'no' was mapped to 0.

Feature Scaling

Feature scaling is a crucial preprocessing step, especially when using machine learning algorithms such as neural networks, k-Nearest Neighbours (KNN), or support vector machines (SVM), which are sensitive to the scale of input data. Without scaling, features with larger magnitudes could disproportionately influence the model during training, leading to suboptimal performance.

In this project, the **StandardScaler** was used to normalise continuous variables, such as **total_day_minutes**, **total_eve_calls**, and **total_night_minutes**. The StandardScaler standardises features by removing the mean and scaling them to unit variance, ensuring that each feature contributes equally during model training. This is particularly important when working with neural networks, where unscaled inputs can slow down the convergence of the model or lead to poor results.

Final Data Preparation

After completing the essential preprocessing steps, the final preparation of the data was carried out to structure it appropriately for model training and evaluation. The goal was to split the dataset into training and testing sets and ensure that all features and labels were correctly aligned.

Splitting the Data into Features and Target

The input features for the training set were extracted from the preprocessed data by selecting all columns starting from the third index onward (excluding the ID and other non-relevant columns). The target variable (**churn**) was assigned to the corresponding labels.

Train-Test Split

To evaluate the model's performance effectively, the dataset was split into training and testing sets. This step ensures that the model is trained on one portion of the data while being tested on another unseen portion to assess its generalisation performance. The **train_test_split** function from scikit-learn was used to perform this split. The random_state parameter was set to 0 to ensure reproducibility of the results.

This final preparation step ensures that the training and testing data are properly organised for subsequent model development. The training data was then used to fit the models, while the test data was reserved to validate their performance.

Model Development

In this project, two predictive models were developed to address the task of customer churn prediction: **K-Nearest Neighbours (KNN)** and an **Artificial Neural Network (ANN)**. Both models were chosen for their effectiveness in solving classification problems, particularly binary classification tasks like churn prediction. This section discusses the architecture, rationale, and performance of both models.

K-Nearest Neighbours

Model Overview

The **K-Nearest Neighbours (KNN)** algorithm is a simple yet powerful method often used for classification tasks. It operates by finding the ‘k’ nearest data points (neighbours) to a given observation and assigning a class based on the majority class of those neighbours. KNN works particularly well for classification tasks where the underlying assumption is that similar data points are likely to belong to the same class.

For this project, **k** was set to **4**, meaning that for each customer, the model looked at the 4 closest customers (in terms of feature similarity) and classified the customer as churn or not churn based on the majority class of these neighbours. The KNN model was implemented using the scikit-learn library.

Rationale for KNN

The choice to use KNN for the churn prediction task was motivated by the model's simplicity and interpretability. Since the dataset consisted of several continuous variables such as total call duration and customer interaction data, KNN was suitable due to its ability to make predictions based on feature similarity. Additionally, KNN does not require a complex mathematical model or assumptions about the distribution of data, making it a versatile and flexible approach.

The simplicity of KNN allowed for rapid experimentation and tuning. With the number of neighbours set to 4, the model was able to capture the key relationships in the dataset, leveraging the natural clustering of customers based on their behaviour. The use of the StandardScaler during preprocessing ensured that all features contributed equally to the distance calculations, further improving the performance of KNN.

Model Performance

The KNN model was trained on the processed data, and the results were highly satisfactory. It achieved an accuracy of over **90%** on the test data, indicating that the model was able to accurately identify the majority of customers who would churn. The performance was considered strong, given the simplicity of the model and the relatively low computational overhead required.

One of the main advantages of KNN is its interpretability. By looking at the neighbours of a data point, it is possible to directly observe why a specific classification was made, allowing for clearer insights into the decision-making process. This feature can be useful in customer churn prediction, where understanding the reasoning behind churn predictions is often as valuable as the prediction itself.

Artificial Neural Network

Model Overview

The second model implemented was an **Artificial Neural Network (ANN)**. ANNs are a class of machine learning models inspired by the human brain and are particularly effective for complex classification problems involving non-linear relationships between features. ANNs consist of layers of interconnected nodes (neurons), where each node applies a transformation to the input data and passes it on to the next layer. Through multiple layers, ANNs can model complex patterns in data.

For this project, an ANN was developed with the following architecture:

- **Input Layer**: Consisting of the scaled features after preprocessing.
- **Two Hidden Layers**: Each hidden layer contained **6 units**, and the **ReLU (Rectified Linear Unit)** activation function was used. ReLU was chosen due to its ability to introduce non-linearity into the model while avoiding issues like vanishing gradients.
- **Output Layer**: The output layer had one unit with a **sigmoid activation function**. This function is commonly used for binary classification tasks, as it outputs a probability between 0 and 1, which can be interpreted as the likelihood of a customer churning.

The model was trained for **100 epochs** with **binary cross-entropy** as the loss function and **Adam optimiser** for weight updates. The dataset was split into training and test sets, with the model being evaluated on unseen test data after training.

Rationale for ANN

The rationale for implementing an ANN was driven by the need to capture more complex, non-linear relationships between the features in the dataset. Telecom customer churn is influenced by a variety of factors, some of which may not have simple linear relationships with the target variable. An ANN, with its multi-layer architecture and ability to model non-linear patterns, was well-suited to address this complexity.

Additionally, ANNs can handle large datasets with many features and automatically learn important representations without the need for manual feature engineering. Given the diverse set of features in the telecom dataset, ranging from call durations to customer service interactions, the ANN provided a flexible approach to capturing these patterns effectively.

Model Performance

After training the ANN for 100 epochs, the model achieved an impressive accuracy of **93%** on the test data. This result demonstrated that the ANN was capable of making highly accurate predictions about customer churn. The high performance of the ANN can be attributed to its ability to model the non-linear interactions between the features, particularly those related to customer service interactions and usage behaviour.

One of the challenges associated with ANNs is their "black-box" nature, where the decision-making process is not as easily interpretable as simpler models like KNN. However, this trade-off was justified by the significant improvement in accuracy. Moreover, the ANN's ability to learn complex representations made it a powerful tool for understanding the intricate patterns driving customer churn.

Conclusion on Model Development

In conclusion, both the KNN and ANN models performed well in predicting customer churn, with each offering unique advantages. The KNN model provided simplicity and interpretability with a strong accuracy of over 90%, while the ANN delivered superior performance with a 93% accuracy, leveraging its ability to model complex non-linear relationships. The combination of these two approaches provided a robust solution for the task of predicting customer churn in the telecom sector.

Both models were tested on unseen data, and their predictions can be used to guide business decisions such as identifying at-risk customers and implementing targeted retention strategies. The balance between simplicity and complexity offered by these models makes them highly effective for solving real-world churn prediction problems.

Model Evaluation

The performance of both the KNN and ANN models was primarily evaluated using **accuracy** as the key metric. Accuracy is a straightforward and effective measure for binary classification problems like customer churn, as it calculates the proportion of correct predictions out of all predictions made.

- **KNN**: The KNN model achieved an accuracy of over **90%** on the test data. Despite being a simpler model, KNN performed strongly in predicting customer churn. Its ability to classify based on similarity among customers contributed to its effectiveness. The model was later converted into a .pkl file for deployment on a **Streamlit** app, demonstrating its utility in real-world applications, even though its accuracy was slightly lower than the ANN model.

- **ANN**: The ANN model outperformed KNN with an accuracy of **93%**, thanks to its ability to model complex relationships within the data. The deeper architecture of the neural network, along with the ReLU and sigmoid

activation functions, allowed it to capture more intricate patterns, particularly those related to customer behaviour and interactions.

To ensure the robustness of the models, **cross-validation** was performed on both KNN and ANN. This method splits the data into different folds and evaluates the model on multiple splits to verify its consistency. Both models showed consistent performance across various data partitions, reinforcing their reliability.

Despite the KNN model's 3% lower accuracy compared to the ANN, its simplicity and ease of deployment made it a valuable tool for showcasing predictions, while the ANN excelled in capturing deeper data patterns. Both models provided effective solutions for customer churn prediction.

Results and Discussion

In this study, both the **K-Nearest Neighbours (KNN)** and **Artificial Neural Network (ANN)** models were developed to predict customer churn in the telecom sector, with the goal of identifying at-risk customers based on their usage patterns and interaction with the service. The performance of these models was compared, and several important insights were derived from the analysis.

The **ANN model** emerged as the stronger predictor, achieving an accuracy of **93%**, outperforming the **KNN model**, which achieved an accuracy of **90%**. The improved accuracy of the ANN model can be attributed to its deeper architecture and ability to capture complex, non-linear relationships between the features. This model effectively leveraged feature scaling, activation functions, and training over 100 epochs to optimise its predictive capabilities. On the other hand, the KNN model, while slightly less accurate, performed admirably given its simplicity and lack of extensive parameter tuning.

Several key factors were identified as having a significant influence on customer churn:

- **Customer Service Calls**: One of the most important predictors of churn was the number of customer service calls made by the customers. The data revealed that customers who had frequent interactions with customer service were more likely to churn. This could indicate that these customers were experiencing unresolved issues or dissatisfaction with the service. Both the KNN and ANN models effectively captured this relationship, recognizing that a high number of customer service calls is often a red flag for potential churn.

- **Call Duration**: Another critical factor was the total call duration, particularly during the day. The data showed that customers who had higher call durations during the daytime were less likely to churn. This finding suggests that customers who are more engaged with the service, as measured by their call usage, are generally more loyal. These high-usage customers likely derive more value from the service, reducing their propensity to switch providers. Both models picked up on this trend, with the ANN model being

slightly more precise in distinguishing between high- and low-engagement customers.

- **International Plan**: The presence of an international plan was also a significant factor in predicting churn. Customers who had subscribed to an international plan exhibited different usage behaviours compared to those without the plan and were generally less likely to churn. This could be due to the perceived value of the plan or the increased complexity involved in switching providers when such specialised services are in use. The ANN model, with its ability to capture these more intricate patterns, outperformed the KNN model in predicting churn among customers with international plans.

The **KNN model**, although simpler, remained a valuable tool for churn prediction. With a k-value of 4, the KNN model classified customers based on their similarity to the closest four neighbours, effectively identifying patterns related to churn. The primary advantage of KNN lies in its interpretability and the ease with which it can be applied to classification tasks. While its performance was slightly lower than that of the ANN model, its simpler design and transparent decision-making process make it suitable for applications where computational efficiency and ease of deployment are prioritised.

The **ANN model**, however, demonstrated a superior ability to handle complex patterns and interactions within the data. Its architecture, consisting of two hidden layers with 6 units each and ReLU activation functions, allowed it to capture non-linear relationships that the KNN model might miss. The use of a sigmoid activation function in the output layer enabled precise binary classification, which is particularly useful for distinguishing between churn and non-churn customers. The model's accuracy of 93% reflects its capability to generalise well to unseen data and its robustness in identifying churn patterns across different customer groups.

In addition to these key findings, both models demonstrated robustness through cross-validation, where the data was split into different subsets to evaluate the models' consistency. Both the KNN and ANN models showed stable performance across various data splits, further validating their effectiveness as predictive tools for customer churn. The ANN model, in particular, displayed minimal variance across splits, further reinforcing its ability to generalise to different subsets of data.

In conclusion, the results of this analysis highlight the **ANN model** as the more powerful and accurate tool for predicting customer churn in the telecom sector, especially when it comes to capturing complex behavioural patterns. However, the **KNN model** remains a valuable and interpretable alternative, particularly for cases where simplicity and ease of implementation are priorities. Both models provide useful insights into the factors driving customer churn, enabling telecom companies to proactively identify at-risk customers and implement retention strategies.

Conclusion

This project successfully developed and evaluated two predictive models—**K-Nearest Neighbours (KNN)** and an **Artificial Neural Network (ANN)**—to address the issue of customer churn in the telecom sector. Both models offered valuable insights into customer behaviours, helping identify those at risk of leaving the service. The **ANN model**, with an accuracy of **93%**, outperformed the **KNN model** (which achieved over **90% accuracy**), demonstrating its superior ability to capture complex patterns and relationships within the data. The deep learning architecture of the ANN allowed it to generalise more effectively and offer more precise churn predictions.

The project uncovered key factors influencing customer churn, such as the number of customer service interactions, call duration during the day, and whether or not a customer had subscribed to an international plan. These insights are crucial for telecom companies aiming to develop data-driven, targeted retention strategies. For instance, the strong correlation between customer service calls and churn suggests that improving customer support processes could reduce churn rates. Similarly, understanding how high engagement during the day translates to lower churn rates can guide strategies to encourage increased usage among customers.

While both models achieved strong performance, the **ANN model** demonstrated a clear advantage in predictive accuracy and robustness. It was able to handle more intricate relationships in the data, particularly when considering the impact of multiple variables on customer churn. Despite this, the **KNN model** proved to be a useful and interpretable tool, offering an alternative approach to churn prediction that is simpler and easier to implement.

The project has successfully demonstrated how machine learning models can be used to predict customer churn, offering actionable insights that can help telecom companies retain more customers. By identifying patterns and predictors of churn, these models enable proactive interventions to prevent at-risk customers from leaving.

Looking ahead, there is considerable potential to further improve the prediction performance and the depth of insights gained from the models. **Future work** could explore more advanced machine learning algorithms, such as **Random Forests** or **Gradient Boosting**, which may offer even greater predictive power. Additionally, further **feature engineering could** enhance the models by identifying new variables or combining existing features in ways that better capture the nuances of customer behaviour. Another promising avenue is to incorporate **time-series data**, analysing how customer interactions evolve over time, which could provide even deeper insights into the drivers of churn. Such an approach could reveal trends in customer engagement that static features might miss, leading to more sophisticated retention strategies.

In conclusion, the project not only delivered high-performing predictive models but also offered critical insights into the factors that drive customer churn. These findings can serve as a foundation for telecom companies looking to enhance

customer retention and reduce churn, providing them with the tools and strategies needed to maintain a loyal customer base.

References

This project utilised a variety of tools, libraries, and resources to build and evaluate the predictive models for customer churn. The following references were instrumental in the successful completion of the project:

1. **Libraries and Tools:**
    - **Pandas**: Used for data manipulation and preprocessing.
    - **NumPy**: Assisted in numerical computations.
    - **Scikit - learn**: Utilised for implementing the KNN algorithm, train-test splitting, cross-validation, and evaluation metrics such as accuracy.
    - **TensorFlow/ Keras**: Leveraged for building and training the Artificial Neural Network (ANN) model, providing efficient implementation of neural network layers and optimisation functions.
    - **Matplotlib**: Used for visualising the data during Exploratory Data Analysis (EDA), including histograms, scatter.
2. **Model Deployment:**
    - **Streamlit**: Enabled the deployment of the KNN model, allowing it to be showcased through an interactive web app.
3. **Research and Documentation:**
    - Various **online articles, tutorials, and official documentation** were referenced throughout the project for understanding best practices in model development, data preprocessing, and evaluation techniques.

THANK YOU SKILL VERTEX.