Assessment 2: Predicting Customer Churn in the Telecommunications Industry Using Machine Learning Models

Valeria Roman Restrepo **– ID.** 24896716

Course: 32513 31005 Advanced Data Analytics Algorithms, Machine Learning

Spring 2024

UTS

Table of Contents

1). Business Understanding

1.1). Background and context

Maintaining customer retention is getting more challenging than ever in today's fast-paced world. The churn rate of telephone service companies has been increasing dramatically, which is why they want to attract more and more subscribers and prevent existing customers from terminating their contracts and moving on to another company (churn rate). For a telephone company to continue to expand its customer base, its growth rate (number of new customers) must exceed its churn rate (number of existing customers). Some factors that cause existing consumers to leave such companies are better prices, faster internet services and a more secure online experience than other companies (Kumar, S. 2020).

Telephone companies routinely collect large amounts of customer data, encompassing demographic information, usage patterns, payment history and customer service interactions. This research project focuses on leveraging this rich dataset to build a predictive model that identifies customers at risk of dropping out of service using three machine learning models: Logistic Regression, Catboost and Random Forest. The aim is to compare their best performance to achieve a system that detects the risk of losing customers early. This way, this company's marketing team can proactively develop campaigns to retain clients, offering better promotions and loyalty rewards or attacking problems related to the services.

The dataset used for the proper resolution of this study includes customer demographic information such as gender, whether the customer is a senior citizen or not and whether the customer has dependents. On the other hand, it also contains specific information related to the services they consume. Furthermore, it captures financial information such as monthly and total charges, as well as the customer's method of payment (Blastchar, nd). With this comprehensive dataset, the aim is to pinpoint critical patterns and variables that drive the majority of customer churn, thereby enabling the company to design key retention strategies that can be the cornerstone for business growth. In this way, the company can improve customer satisfaction, reduce churn, and ultimately enhance profitability.

1.2). Objectives

The project aims to create a Machine Learning predictive model for estimating churn (customer turnover or 'customer attrition'). The main objective is to accurately and effectively predict which existing customers (subscribers) will likely quit Telecom by analysing a dataset of customers' demographic and service usage information, including their service packages, payment transaction history and support call logs.

Three goals were unfolded to achieve the general goal. The first was developing detailed feature engineering/feature preprocessing that was most suitable for the model. The project's second goal will be featuring selection by pinpointing the most relevant features to customer churn. It will also develop a thoroughly detailed feature engineering to improve the model's performance if it misses any desired features (sometimes, various human-sounding features cannot be entirely extracted and can be misleading to the model). The third goal was to develop machine learning models and fine-tune their model parameters (logistic regression, CatBoost, and

Random Forest) before utilising them to achieve the best performance. We would like to predict the churners (leavers) from the telecom company.

Ultimately, the model's outputs will be translated into strategic actions for the business: Running campaign testing drawn from specific market segments and developing new protocols for improving customer service.

1.3). Impact of the project

By implementing a machine learning model, the telecom company can accurately predict which customers are likely to leave; this will help reduce churn by allowing the company to focus on high-risk customers before they decide to leave. Since retaining existing customers is more cost-effective than acquiring new ones, this approach will form the foundation for a more profitable business.

When even a few customers leave, their lifetime value can increase; this might sound counterintuitive, but it is an essential factor for the company as it works to get the most out of its existing customer base. This project aims to help the company make more thoughtful use of its resources by focusing on customers most likely to leave. While broad strategies can be helpful, targeted efforts—like offering personalized discounts or improving customer service—usually have a much more significant impact. By using insights from the predictive model, these actions can boost customer satisfaction, help cut costs and make things run more efficiently for the company.

This research project will help the telecom company build stronger customer relationships and create strategies to keep them loyal. It will also play a significant role in shaping a more effective loyalty program. By reducing customer churn, the company can ensure steady profits over the long term, with the predictive model helping to make more informed, customer-driven decisions. This approach will give the company a natural edge in today's competitive market.

2). Data Understanding

The dataset used for the resolution of this research project is available at Kaggle (Blastchar. n.d.). It is about a company that supplies home and internet services to 7043 customers in California, which stems from the IBM sample set collection. The target variable is Churn, a binary field indicating whether a customer has left the company (Yes) or stayed (No).

The other features exposed in the dataset can be grouped into different categories, each adding value to the predictive model:

1). Customer demographic information, including gender, age, and marital status.

2). Customer account information includes the number of months customers have been with the company, paperless billing, payment method, and monthly and total charges.

3). Customer usage behaviour, such as TV and movie streaming services.

4). Services to which the customers signed up, such as phone, multiples, internet, online insurance, online backup, device protection, and technical support.

5). Customer churn, where we can see if a customer has left the company in the last month (Kumar, S. 2020).

The above can be seen in the data dictionary in **Figure 1** in Appendix.

3). Data Preparation

The data preparation involved seven steps that can be seen in **Figure 2** and are described below.

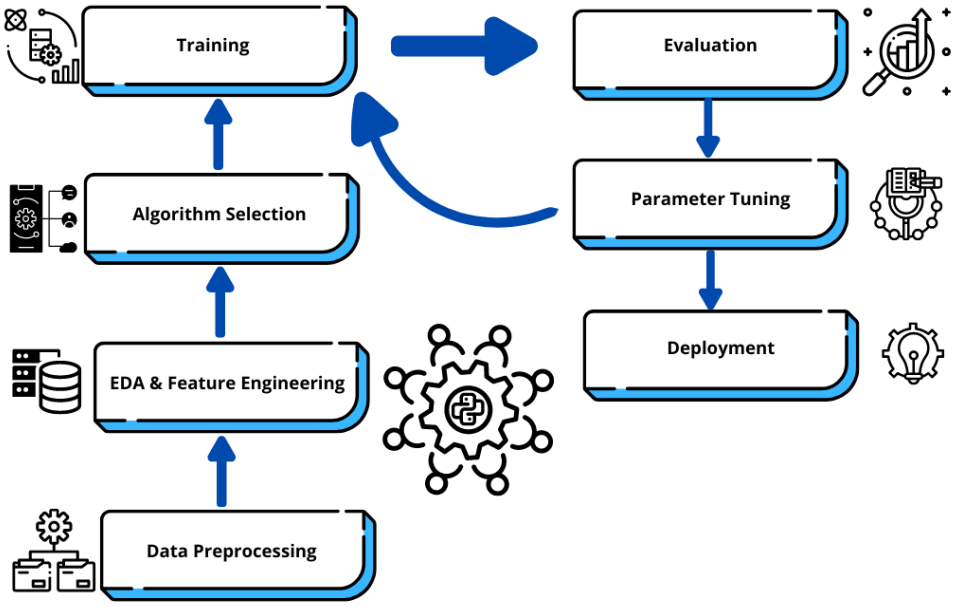

Figure 2. *Steps for data Preparation.*

- **Data Preprocessing**

  We started by eliminating the customerID column, as it is a single indicator and does not provide the model with value-added information.

  Then, the df.describe() function was used to help us make decisions in the preprocessing stage, such as handling missing values, normalising or standardising the data, and dealing with outliers. These steps are crucial when preparing the data effectively for applying machine learning models. Those results can be seen in **Figure 3**.

	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

Figure 3. *Output of df.describe().*

We then investigated the data types to ensure that the preprocessing was effective and accurate, and this helped us to know which steps to follow and ensure compatibility with the different machine learning models. In our case, we have one float variable, two int and seventeen objects, and based on this, we made the necessary transformations for the columns that did not have the correct data type.

Subsequently, the quality of the data was monitored, as it is a critical step in this project; it helped us to ensure that the dataset was accurate, complete and consistent, which is essential to build reliable models that can deliver meaningful results that are per the objective proposed in the initial phase of this project. By performing this step, we avoided biased predictions by ensuring completeness at every step.

Correlated to the previous check for null values, it could be noted that only the TotalCharges column has a small number of null values with a percentage of 15% as we can see in **Figure 4.** Therefore, they could be eliminated from the dataset. When checking duplicates, it was possible to see that we only had 22 repeated rows, so they could be removed to ensure that the data was clean, prevent potential biases, reduce the risk of overfitting and build a path for a more efficient model training process. As for outliers, none were found.

	Null_Counts	Percentage
gender	0	0.000000
SeniorCitizen	0	0.000000
Partner	0	0.000000
Dependents	0	0.000000
tenure	0	0.000000
PhoneService	0	0.000000
MultipleLines	0	0.000000
InternetService	0	0.000000
OnlineSecurity	0	0.000000
OnlineBackup	0	0.000000
DeviceProtection	0	0.000000
TechSupport	0	0.000000
StreamingTV	0	0.000000
StreamingMovies	0	0.000000
Contract	0	0.000000
PaperlessBilling	0	0.000000
PaymentMethod	0	0.000000
MonthlyCharges	0	0.000000
TotalCharges	11	0.156183
Churn	0	0.000000

Figure 4. *Null count analysis.*

EDA & Feature Engineering

Exploratory data analysis (EDA) is a pivotal step in understanding the patterns and relationships within the dataset, which aids in feature engineering and model development. In this research project, EDA unveiled relevant findings such as the skewed distribution of SeniorCitizen, the high proportion of low-tenure customers and the bimodal distribution of MonthlyCharges, all suggesting influential factors in predicting the target variable. Additionally, the long tail in TotalCharges indicates that tenure is more correlated with this variable than with the others, and this helps us to discern possible trends for the prediction, as we can appreciate in **Figure 5.**

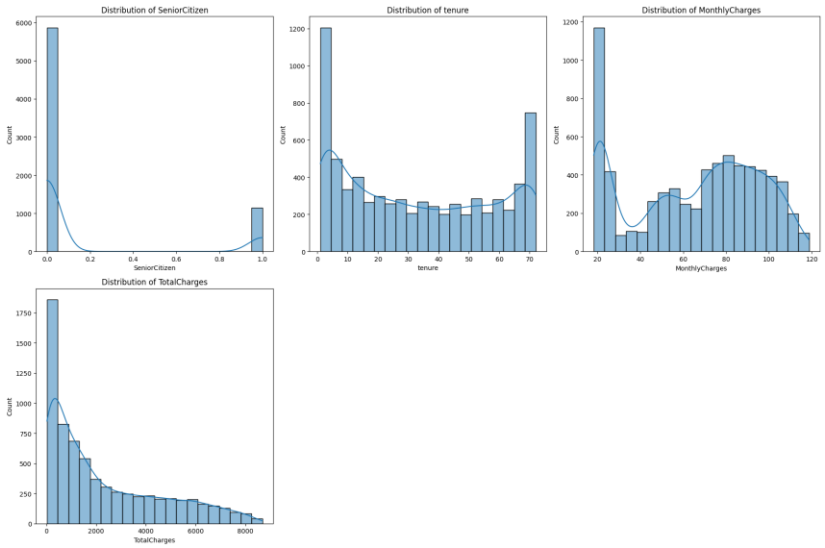

Figure 5. *Distribution of the numerical variables.*

To complement the analysis of the distribution mentioned earlier, a correlation heatmap was developed to quantify the relationships between the variables SeniorCitizen, tenure, MonthlyCharges, and TotalCharges. The heatmap in **Figure 6** shows a strong positive correlation (0.83) between tenure and TotalCharges; this is consistent with the previous observation of the long tail in the distribution for TotalCharges, implying that customers who stay longer generate more total charges than the ones who leave the company earlier. From this, it was determined that the TotalCharges column could be eliminated as it is highly correlated with tenure and MonthlyCharges (correlations of 0.83 and 0.65, respectively).

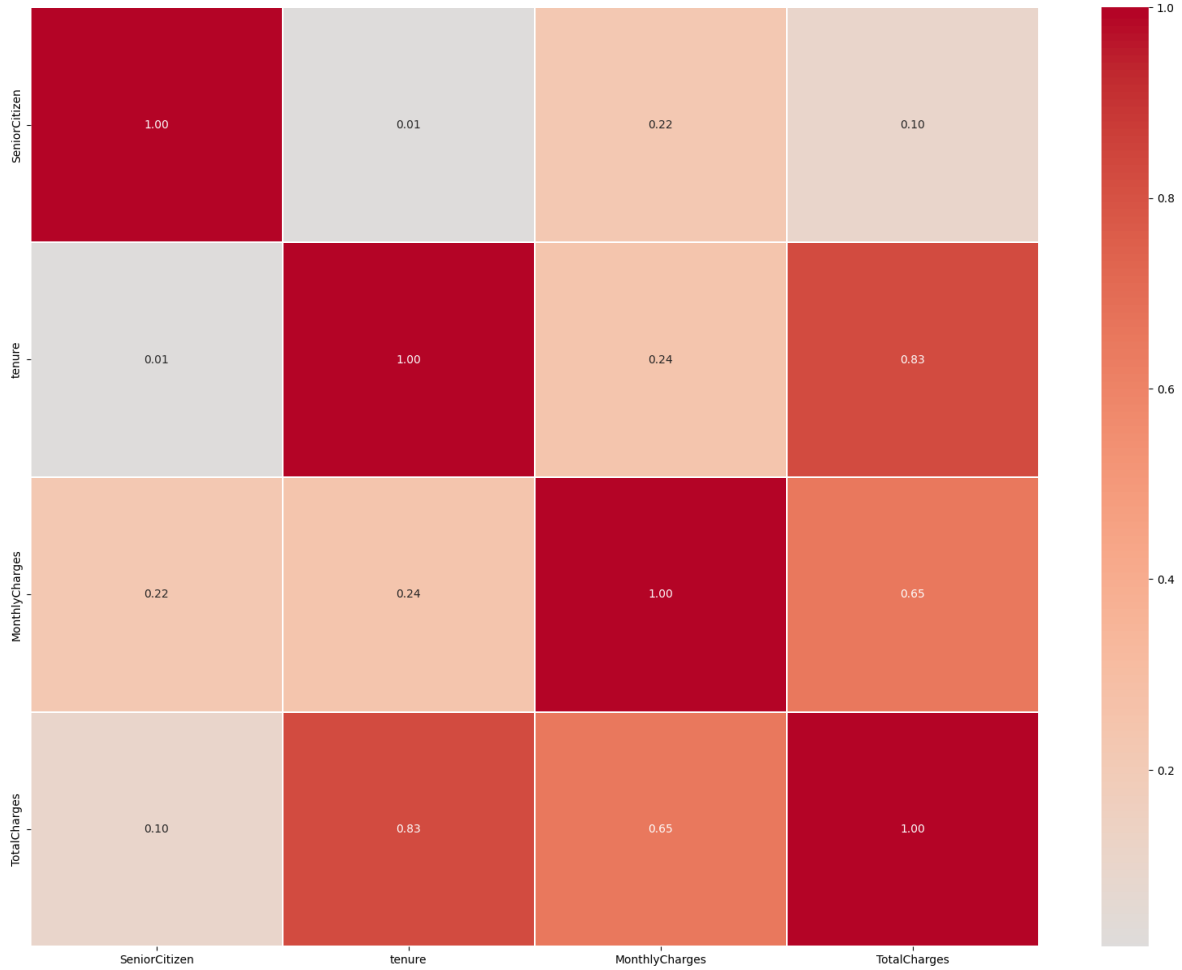

Figure 6. *Heat map.*

A pair plot was created to corroborate these numerical variables' importance in predicting Churn, as shown in **Figure 7.** This visual analysis suggests that tenure and monthly charges are indeed features of critical interest and should be retained in contrast to total charges, which can be eliminated as was done in the previous step.

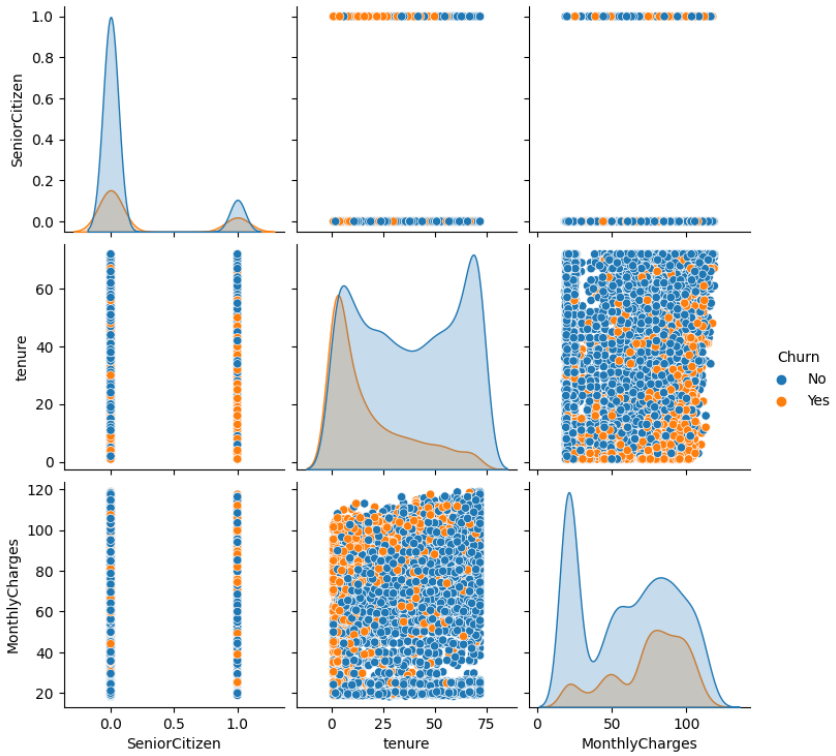

Figure 7. *Pair plot.*

When analysing categorical variables about churn, it is essential to highlight the influence of customer behaviour. Variables such as contract type, payment method, and whether customers use online security or technical support are essential in predicting churn. Delving deeper into the topic, we can see that customers with month-to-month contracts are likelier to leave the company than those who have been with the company for extended periods. Similarly, customers who need more certainty in the security of the services or who use a specific payment method may exhibit different churn patterns; this complements the previous analysis of continuous variables such as tenure and Monthly Charges, providing a broader understanding of customer behaviour by incorporating both categorical and numerical insights.

Finally, the distribution of the target class was checked, as shown in Figure 8, which is crucial as it reveals the imbalance between "Churn" and "No Churn". This imbalance significantly impacts the model's performance, especially in classification tasks. If one class (in our case, "No") dominates the dataset, the model may be biased towards predicting this class more frequently, leading to inaccurate results. By evaluating this class distribution, we determined the models that should be applied to address the proposed objective correctly.

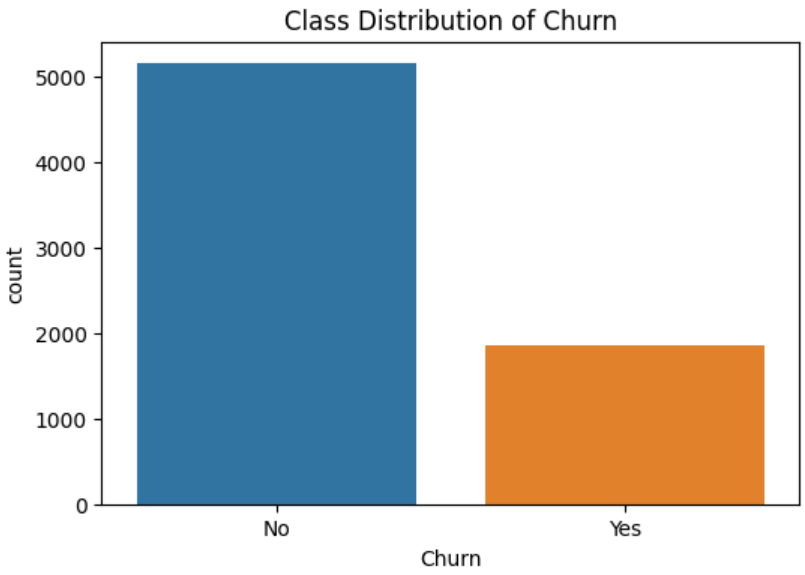

Figure 8. *Class Distribution of Churn.*

- **Algorithm Selection**

  We looked at three machine learning models: Logistic Regression, Random Forest, and CatBoost. Each one has its strengths when it comes to tasks like predicting churn. Logistic Regression is a straightforward, easy-to-understand model that works well but can need help handling more complex, non-linear data. Meanwhile, a Random Forest is a collection of decision trees that work together to avoid overfitting and better handle complex, non-linear data by averaging the results from all the trees. CatBoost, on the other hand, is a gradient-boosting model that shines when working with categorical data, making it perfect for our dataset, especially since it is imbalanced.

- **Training**

  During the training phase, we started with a simple baseline model and the three more advanced machine-learning models we mentioned earlier. The baseline model gave us a good reference point, allowing us to see how much the more sophisticated models improved performance. To train the models, we split the dataset into three parts: one for training, another for validation, and the last for testing. Each model was trained on the first portion, fine-tuned using the validation set, and tested on the final set to see how well it performed. The baseline gave us a solid starting point, while the advanced models were fine-tuned through hyperparameter adjustments to deal with class imbalances and boost accuracy; this structured process made it easy to see how better each model was at predicting customer churn compared to the baseline.

- **Evaluation**

  We measured how well the models performed using accuracy, precision, recall, and the F1 score. These metrics gave us a clear understanding of how each model balanced the risk of false positives and false negatives, which is essential when predicting which customers might leave. Since our dataset was imbalanced, recall was crucial because

our main goal was to spot at-risk customers. We looked closely at how the more advanced models—Logistic Regression, Random Forest, and CatBoost—stacked against the baseline model. We focused on how well each model handled the data imbalance and improved recall. After some adjustments, CatBoost performed better than the others, making it the best choice for predicting which customers might leave.

- **Parameter Tuning**

The description of the chosen hyperparameters for tuning the models is presented in Table 1.

Model	Best parameters for tuning	Explanation
Logistic Regression	C = 1	Regularisation strength is moderate, balancing between overfitting and underfitting.
	max_iter = 200	Solver had enough iterations (200) to converge properly.
	penalty = 'l2'	Ridge regularisation helps in reducing model complexity and preventing overfitting by penalising large coefficients.
	solver = 'saga'	Efficient for large datasets and works well with L2 regularisation.
Catboost	depth = 6	Allows the model to capture non-linear patterns without overfitting, striking a balance between model complexity and performance.
	iterations = 100	Provides a good balance between computational efficiency and model accuracy without overfitting.
	l2_leaf_reg = 1	Regularisation coefficient used to prevent overfitting. A value of 1 indicated a mild regularisation to maintain simplicity and generalise well on unseen data.
	learning_rate = 0.1	Is a common starting point. It helps in gradual learning and avoids drastic updates to the model parameters, ensuring steady improvements in the model's performance.
Random Forest	bootstrap = True	Enables bootstrapping, allowing sampling with replacement for better generalization.
	max_depth = 10	Limits the maximum depth of the trees to prevent overfitting.
	min_samples_leaf = 1	Ensures that each leaf node has at least 1 sample.
	min_samples_split = 2	Minimum number of samples required to split a node.
	n_estimators = 300	The number of trees in the forest, providing model robustness and stability.

Table 1. *Hyperparameters for tuning.*

4). Modelling

The three models were first trained, then the best parameters were calculated to rerun the model, and finally, the ROC curve and the confusion matrix were calculated to complement the analysis.

5). Evaluation

5.1). Results and Analysis

When analysing the performance of the three advanced machine learning models, the results show a significant improvement when the parameters are tuned compared to the baseline models, as seen in **Table 2.** The baseline models for the three algorithms show similar metrics across the training, validation, and testing sets. When optimised, the models improve significantly, particularly Random Forest, when viewed from the perspective of F1-score. Recall values improved across all models, with the CatBoost model exhibiting a notable balance between precision and recall, making it a robust choice for churn prediction.

Model	Data Split	Accuracy	Precision	Recall	F1 Score
Logistic Regression - Baseline	Train	0.7265	0.3632	0.5000	0.4208
	Validation	0.7246	0.3623	0.5000	0.4206
	Test	0.7710	0.3855	0.5000	0.4354
Logistic Regression - Model	Train	0.7969	0.7475	0.7152	0.7276
	Validation	0.8066	0.7605	0.7361	0.7463
	Test	0.8245	0.7532	0.7274	0.7386
Logistic Regression - Best Params	Train	0.7969	0.7475	0.7152	0.7276
	Validation	0.8066	0.7605	0.7361	0.7463
	Test	0.8245	0.7532	0.7274	0.7386
CatBoost - Baseline	Train	0.7265	0.3632	0.5000	0.4208
	Validation	0.7246	0.3623	0.5000	0.4206
	Test	0.7710	0.3855	0.5000	0.4354
CatBoost - Model	Train	0.8277	0.7952	0.7480	0.7656
	Validation	0.8030	0.7591	0.7196	0.7342
	Test	0.8203	0.7479	0.7115	0.7261
CatBoost - Best Params	Train	0.8132	0.7715	0.7345	0.7487
	Validation	0.8030	0.7558	0.7307	0.7410
	Test	0.8295	0.7612	0.7340	0.7457
Random Forest - Baseline	Train	0.7265	0.3632	0.5000	0.4208
	Validation	0.7246	0.3623	0.5000	0.4206
	Test	0.7710	0.3855	0.5000	0.4354
Random Forest - Model	Train	0.8975	0.8805	0.8555	0.8667
	Validation	0.7959	0.7465	0.7187	0.7298
	Test	0.8124	0.7340	0.7141	0.7228
Random Forest - Best Params	Train	0.8968	0.8803	0.8538	0.8656
	Validation	0.8004	0.7534	0.7218	0.7341
	Test	0.8174	0.7418	0.7195	0.7292

Table 2. *Results of the tree models.*

We can discern that CatBoost surpassed the other models by a slightly higher margin in accuracy, precision, and recall, making it the best-performing model overall, as illustrated in **Figure 9.** The model that follows relatively closely is Random Forest in terms of all the metrics, with Logistic Regression also demonstrating competitive performance despite being last.

This analysis is consistent with the evaluation of the previous metrics, where CatBoost achieved the highest accuracy and precision, suggesting that it correctly identifies true positives while minimising false negatives.

Since this project's central and fundamental focus is identifying customers prone to churn, recall is a critical metric, measuring the model's ability to identify actual positive churn cases correctly. Based on **Figure 9** and the previous analyses, CatBoost and Random Forest are ideal candidates for consideration. They will help the company reach the pinnacle of success and retain its services' most significant number of consumers. The slight but notable improvement with hyperparameter tuning, as seen in both CatBoost and Random Forest, demonstrates the value of optimising models to achieve better predictive performance, especially regarding recall, as is the case here.

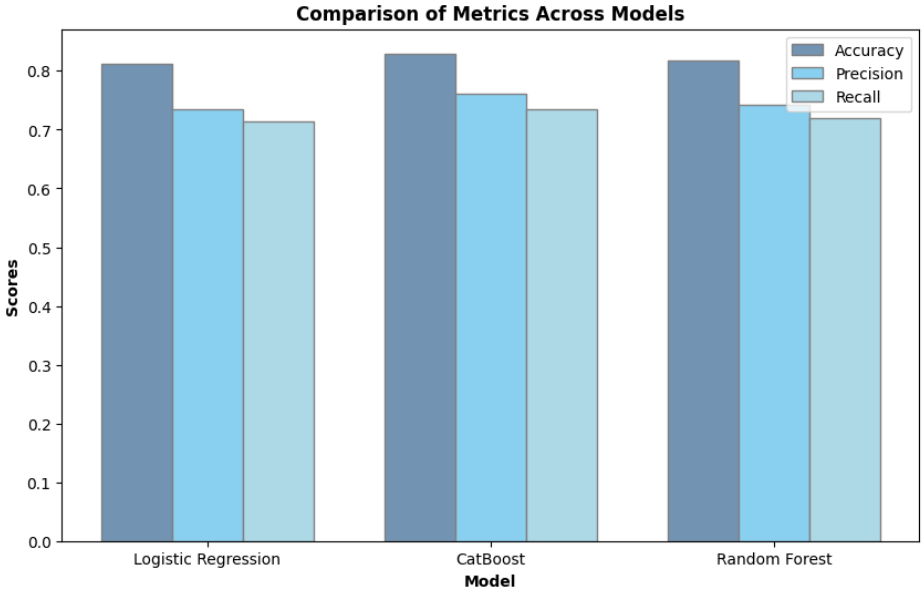

Figure 9. *Comparison of Metric Across Models.*

On the other hand, we can see in **Figure 10** that CatBoost achieves the highest Recall score (0.7340), closely followed by logistic Regression and Random Forest; this aligns with the previously mentioned results where CatBoost also shows better performance in Accuracy and Precision, indicating that it effectively balances the classification of customers who are likely to leave the company with those who are not. Despite being one of the simplest models, Logistic Regression still competes closely with CatBoost regarding Recall performance, demonstrating that it is an effective option for fully meeting this task. Despite having the lowest score in Recall compared to the other models, Random Forest continues to offer competitive performance and can be beneficial in scenarios where high interpretability or feature importance insights are required.

Overall, the Recall scores emphasise that CatBoost is the most effective for identifying customers with the highest risk of churn, which aligns with the business objective of proactively retaining customers.

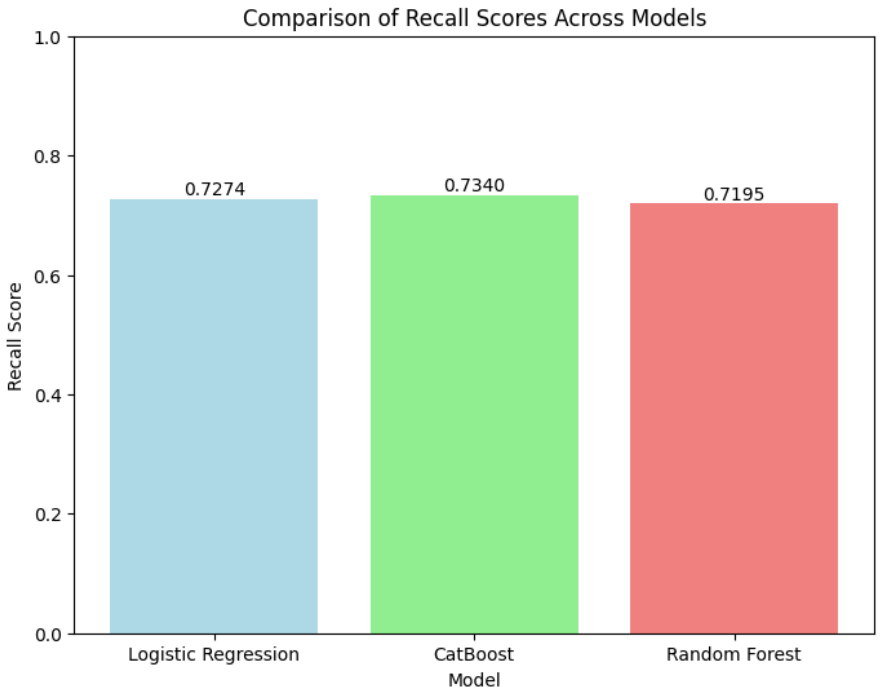

Figure 10. *Comparison of Recall Scores across models.*

Finally, a Streamlit application was created to give the company a visual tool to analyse whether customers are likely to leave the company and to see all the variables that directly influence this decision. Having this tangible tool makes it easier to generate strategic marketing solutions.

5.2). Data Privacy and Ethical Concerns

In this machine learning project, protecting data privacy and addressing ethical concerns is crucial, especially in Australia, where strict laws like the Australian Privacy Act 1988 and the Australian Privacy Principles (APPs) set clear guidelines for handling personal data. Since we are working with sensitive customer data—like demographics and service usage—it is imperative to anonymise the information to protect individual privacy (Australian Government, 1988).

On the ethical side, ensuring the model does not unintentionally discriminate against any group, like senior citizens, could lead to unfair results. We must be transparent about how the model works and explain why specific predictions are made. It is also essential to get explicit consent from customers about how their data will be used, allowing them to opt-out if they are uncomfortable.

6). Conclusion

In conclusion, the CatBoost model was the most effective in predicting customer churn for the telecommunications company, as it achieved the highest recall score; this indicates its superior ability to correctly identify customers at risk of leaving, which is crucial for minimising false negatives and ensuring more targeted retention efforts. The higher performance of CatBoost can significantly assist the company in refining its retention strategies. Additionally, integrating the model into the Streamlit tool provides a user-friendly visual platform, making the insights more accessible and actionable for stakeholders involved in customer retention efforts.

7). References

Australian Government. (1988). **Privacy Act 1988**. Federal Register of Legislation. https://www.legislation.gov.au/Details/C2021C00382

Kumar, S. (2020, October 1). Telco customer churnrate analysis. *Towards Data Science*. https://towardsdatascience.com/telco-customer-churnrate-analysis-d412f208cbbf.

Blastchar. (n.d.). *Telco customer churn* [Data set]. Kaggle. https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

Office of the Australian Information Commissioner (OAIC). (2019). **Australian Privacy Principles Guidelines**. OAIC. https://www.oaic.gov.au/privacy/australian-privacy-principles-guidelines

8). Appendix

Column Name	Description	Data Type
`customerID`	Unique identifier for each customer.	String
`gender`	Gender of the customer (Male, Female).	Categorical
`SeniorCitizen`	Indicates if the customer is a senior citizen (1: Yes, 0: No).	Integer
`Partner`	Whether the customer has a partner (Yes, No).	Categorical
`Dependents`	Whether the customer has dependents (Yes, No).	Categorical
`tenure`	Number of months the customer has stayed with the company.	Integer
`PhoneService`	Whether the customer has phone service (Yes, No).	Categorical
`MultipleLines`	Whether the customer has multiple phone lines (Yes, No, No phone service).	Categorical
`InternetService`	Customer's internet service provider (DSL, Fiber optic, No).	Categorical
`OnlineSecurity`	Whether the customer has online security add-ons (Yes, No, No internet service).	Categorical
`OnlineBackup`	Whether the customer has online backup service (Yes, No, No internet service).	Categorical
`DeviceProtection`	Whether the customer has device protection (Yes, No, No internet service).	Categorical
`TechSupport`	Whether the customer has tech support service (Yes, No, No internet service).	Categorical
`StreamingTV`	Whether the customer has a streaming TV service (Yes, No, No internet service).	Categorical
`StreamingMovies`	Whether the customer has a streaming movie service (Yes, No, No internet service).	Categorical
`Contract`	The contract term of the customer (Month-to-month, One year, Two year).	Categorical
`PaperlessBilling`	Whether the customer has paperless billing (Yes, No).	Categorical
`PaymentMethod`	The payment method the customer uses (Electronic check, Mailed check, Bank transfer, Credit card).	Categorical
`MonthlyCharges`	The monthly amount charged to the customer.	Float
`TotalCharges`	The total amount charged to the customer.	Float
`Churn`	Whether the customer has churned (Yes, No).	Categorical

Figure 1. *Data dictionary.*