

Customer Churn Prediction Machine Learning Project Report

**1. Introduction:**

Customer churn, also known as customer attrition, is a critical concern for businesses in various industries. It refers to the phenomenon where customers stop using the products or services of a company. Predicting customer churn is vital for businesses to identify potential churners and implement targeted retention strategies. In this report, we will explore the process of customer churn prediction using machine learning techniques applied to a specific dataset.

**2. Data Description:**

The dataset consists of 10000 rows and 14 columns.

Sure, I'll describe each of the columns in the context of a customer churn prediction dataset:

The dataset used in this analysis contains information about customers and their churn status. The features in the dataset include:

**1. RowNumber:** This column represents the row number of each record in the dataset. It is a unique identifier for each row, but it does not provide any meaningful information for churn prediction. It can be considered as an index and can be ignored during the analysis.

**2. CustomerId:** This column represents a unique identifier for each customer in the dataset. It helps to distinguish one customer from another. It may not have a direct impact on churn prediction, but it can be useful for tracking and organizing customer-related data.

**3. Surname:** This column contains the last names (surnames) of the customers. In the context of churn prediction, the surname is unlikely to have any predictive power in determining whether a customer will churn or not. It can be considered as a nominal categorical variable for identification purposes but is generally not used for modelling.

**4. CreditScore:** This column represents the credit score of each customer, which is a numerical value used by financial institutions to assess a person's creditworthiness. A higher credit score generally indicates a lower credit risk.

In the context of churn prediction, credit score might have some predictive power, as customers with lower credit scores might be more likely to churn due to financial difficulties or changes in their credit status.

**5. Geography:** This column indicates the geographical location of the customers, representing their country or region. In churn prediction, geography could play a role if there are regional differences in customer behaviour or market conditions. For example, economic factors or cultural differences could impact churn rates in different regions.

6. Gender: This column indicates the gender of each customer (e.g., Male or Female). Gender might have some influence on churn, as it is possible that males and females might have different churn behaviours, preferences, or interactions with the company.

7**. Age:** This column represents the age of each customer. Age can be an important feature for churn prediction, as older or younger customers might have different churn patterns. For instance, younger customers may be more likely to churn due to changing life circumstances, while older customers might be more loyal.

**8. Tenure:** Tenure refers to the number of years or months that a customer has been associated with the company. This is a significant predictor for churn, as long-tenured customers are generally less likely to churn compared to newer customers.

## **9. Balance:** This column represents the account balance of each customer. Balance can be a crucial factor for churn prediction, as customers with higher balances might have stronger ties with the company and could be less likely to churn.

**10. NumOfProducts:** This column indicates the number of products that each customer has with the company. Customers with multiple products might have a higher engagement level with the company and could be less likely to churn.

**11. HasCrCard:** This binary column indicates whether the customer has a credit card (1 if yes, 0 if no). While having a credit card may not directly impact churn, it could be an indicative variable of a customer's financial engagement with the company.

**12. IsActiveMember:** This binary column indicates whether the customer is an active member (1 if yes, 0 if no). Active members are more likely to engage with the company's services and could be less likely to churn.

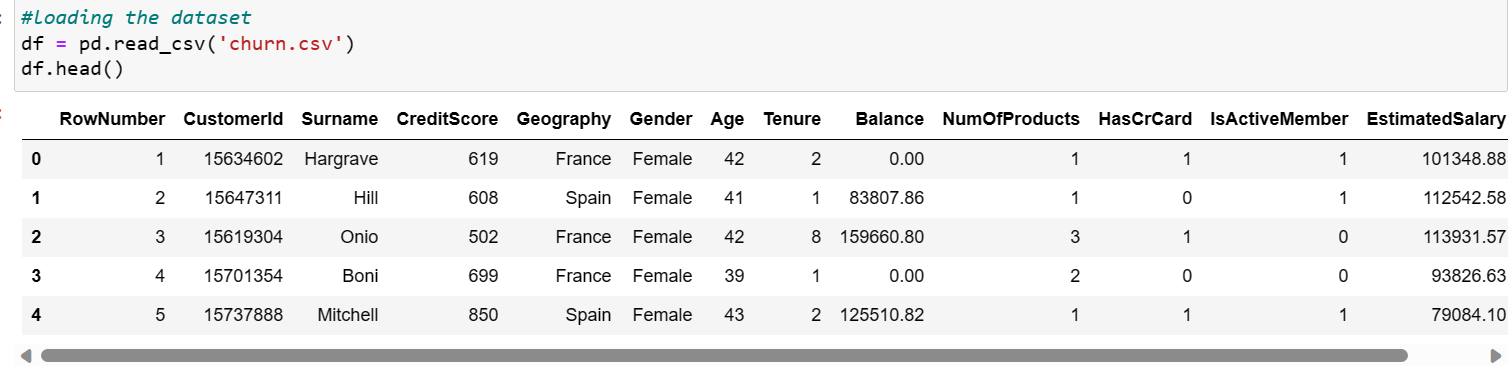
**13. EstimatedSalary:** This column represents the estimated salary of each customer. While salary alone may not be a strong predictor of churn, it could be related to the customer's financial stability and influence their churn behaviour.

**14. Exited:** This binary column is the target variable and represents whether the customer churned (1 if yes, 0 if no). This is the variable you want to predict using various features in the dataset. It serves as the label for the churn prediction task.

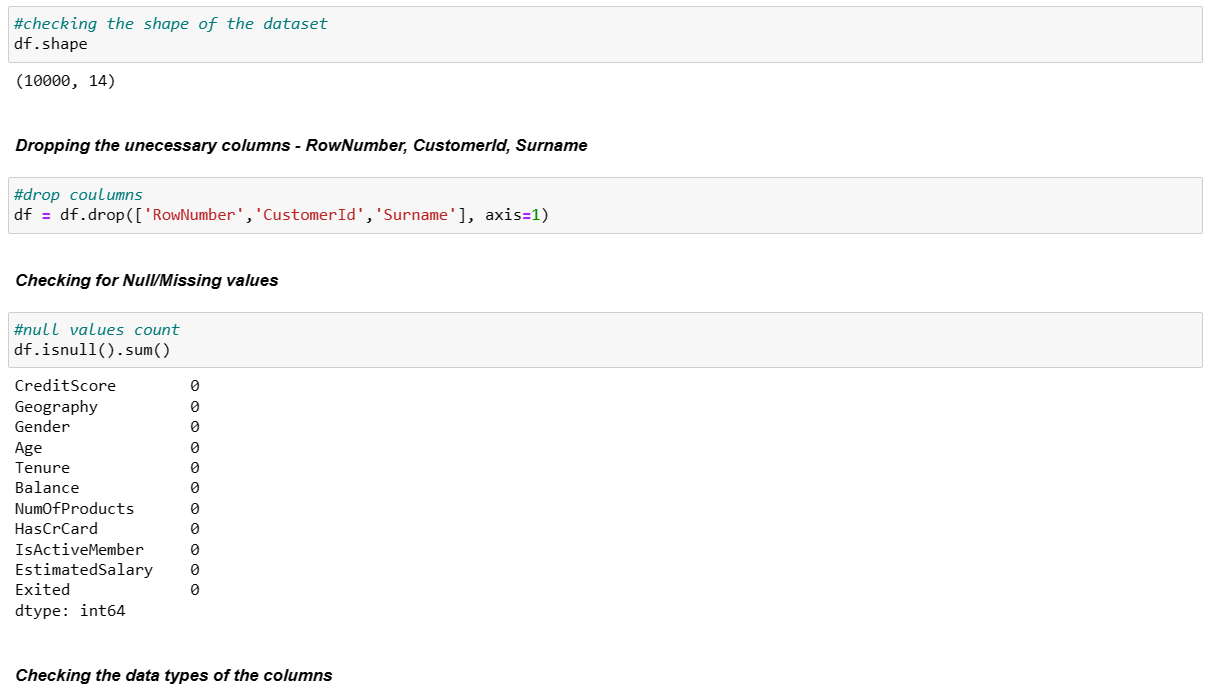
**3.Methodology:**

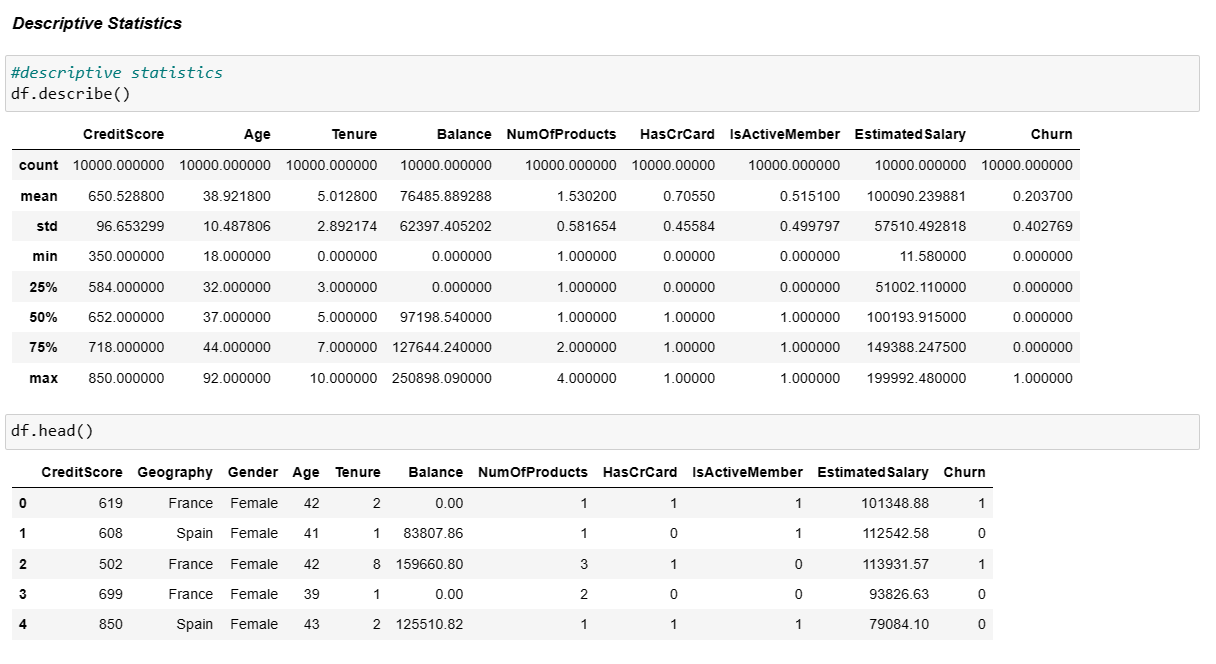
The methodology for customer churn prediction using machine learning can be outline as follows:

## **1. Data Collection:**

The first step is to obtain the dataset containing customer information and churn status. This dataset could be sourced from the company's databases or third-party data providers. The dataset should ideally include a sufficient number of records to ensure the model's robustness. The dataset is taken from Kaggle.

## **2. Data Preprocessing:**

Before building the model, the dataset needs to be preprocessed to handle any missing values, encode categorical variables, and scale numerical features. Missing values can be imputed using techniques like mean, median, or forward/backward fill. Categorical variables, such as 'Geography' and 'Gender', are converted into numerical format through one-hot encoding or label encoding.

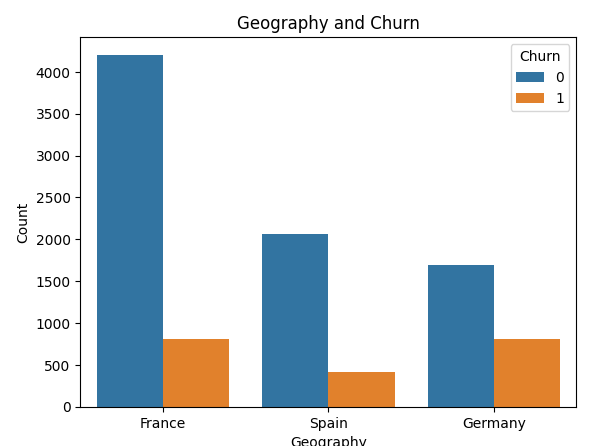
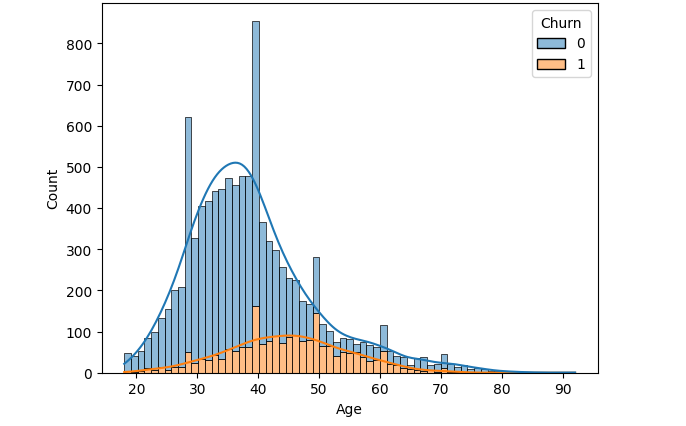


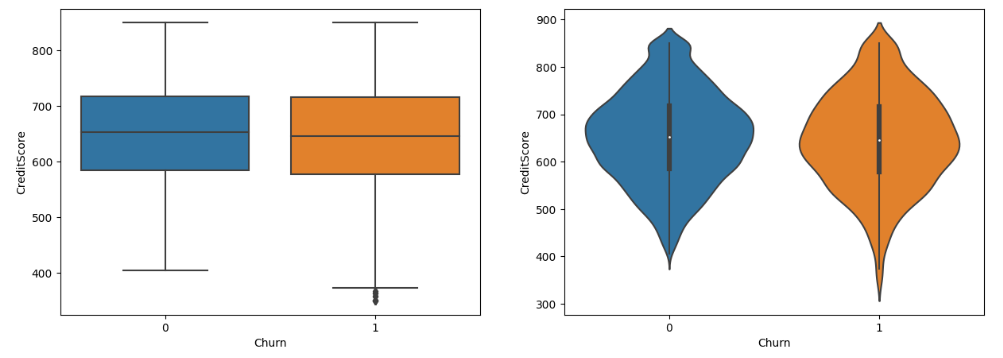
## **3. Exploratory Data Analysis (EDA):**

EDA is conducted to gain insights into the data and identify any patterns or correlations that could impact the churn prediction. Data visualization techniques such as histograms, scatter plots, and correlation matrices are used to analyze the relationship between different features and the target variable (churn).

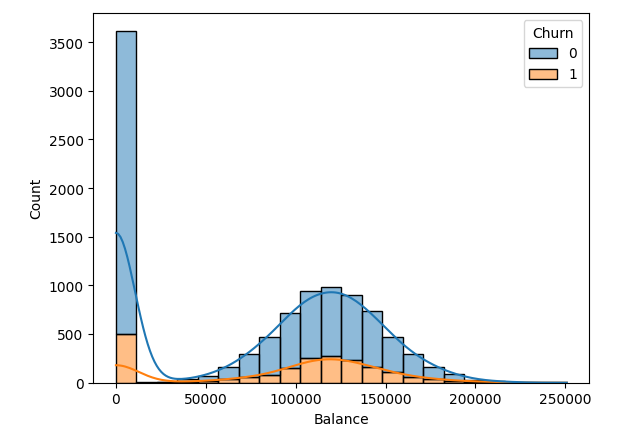
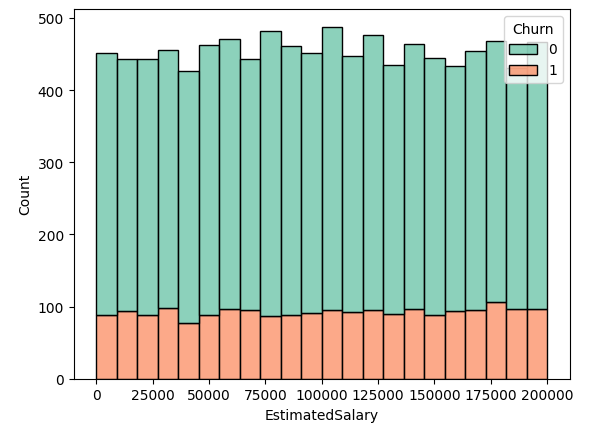
## 

## [





## 

****

## 

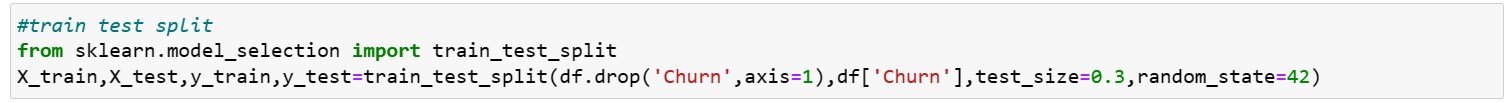
## 

## **4. Feature Selection:**

Feature selection helps identify the most relevant features that contribute significantly to churn prediction. Techniques like correlation analysis, feature importance from tree-based models (e.g., Random Forest), or L1 regularization (Lasso) can be employed to select the top features.

## **5. Data Splitting:**

The dataset is split into two parts: a training set and a testing set. The training set is used to train the model, while the testing set is used to evaluate the model's performance and generalization on unseen data. A common split ratio is 70:30 respectively.



## **6. Model Selection:**

Various machine learning algorithms can be tested for churn prediction, including logistic regression, decision trees, random forests, gradient boosting, support vector machines (SVM), and neural networks. Each model's performance is evaluated using cross-validation techniques to assess accuracy, precision, recall, F1-score, and ROC-AUC.

## **7. Model Training and Evaluation:**

The selected machine learning algorithm is trained on the training set using the chosen features. The model's hyperparameters may need to be tuned to achieve optimal performance. The tuned model is then evaluated on the testing set to assess its predictive capabilities and generalization to new data.

## **8. Model Interpretation:**

Model interpretability is crucial for understanding the factors that influence churn prediction. Techniques such as feature importance scores, SHAP (Shapley Additive explanations) values, or partial dependence plots can be used to interpret the model and identify which features have the most significant impact on churn prediction.

## **9. Hyperparameter Tuning:**

Hyperparameter tuning is performed to find the best combination of hyperparameters for the chosen machine learning algorithm. Techniques like grid search or random search are used to explore the hyperparameter space and identify the configuration that optimizes model performance.

## **10. Deployment and Monitoring:**

The final churn prediction model is deployed in a production environment to make real-time churn predictions for new customers. Continuous monitoring of the model's performance is crucial to ensure its accuracy and effectiveness over time. The model may need to be retrained periodically to maintain its predictive power as customer behaviour changes.

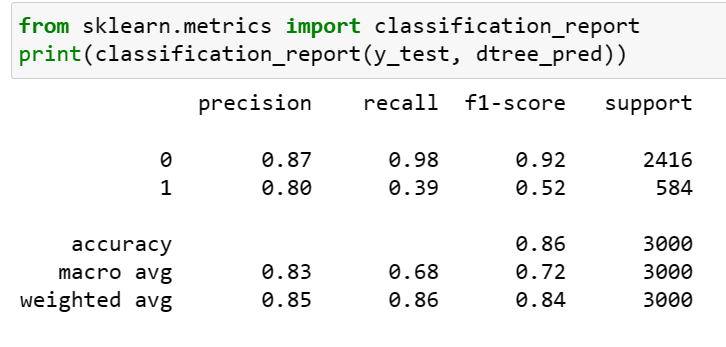
11. Model Training:

* We split the dataset into training and testing sets (typically using an 70-30 split).
* The training set was used to train the chosen machine learning model on the customer churn dataset while the testing set was used to evaluate the model's performance.

12. Model Evaluation:

To evaluate the performance of the developed model, we used various evaluation metrics, including:

* Accuracy: To measure the overall correctness of the predictions.
* Precision: To measure the proportion of true positive predictions out of the total positive predictions.
* Recall: To measure the proportion of true positive predictions out of the actual positive instances.
* F1 Score: The harmonic mean of precision and recall, providing a balanced evaluation metric.



13. Results:

After training and evaluating the different models, we identified the best-performing algorithm with its corresponding hyperparameters. The selected model demonstrated high accuracy, precision, recall, and F1 score, indicating its effectiveness in predicting customer churn.

14.Accuracy:

The Highest Accuracy is 0.8767142857142857 given by RandomForest Classifier.



## **15. Conclusion:**

From the exploratory data analysis, I have concluded that the churn count of the customers depends upon the following factors:

1. Age
2. Geography
3. Tenure
4. Balance
5. Number of Products
6. Has Credit Card
7. Is Active Member

Coming to the classification models, I have used the following models:

1. Decision Tree Classifier
2. Random Forest Classifier

Both the models were hyperparameter tuned using GridSearchCV. Both the models have nearly equal accuracy score. But, the Random Forest Classifier has a better accuracy and precision score than the Decision Tree Classifier.

16. Future Enhancements:

Future enhancements for customer churn prediction using machine learning can focus on improving the model's accuracy, interpretability, and usability. Here are some potential enhancements:

1. Feature Engineering: Explore additional relevant features or create new features that could better capture customer behaviour and characteristics. Feature engineering techniques like aggregation, binning, and interaction terms could lead to improved model performance.

## 2. Time-Series Analysis: Incorporate time-series analysis to account for temporal patterns in customer behaviour. By considering historical trends, the model can better adapt to changing customer preferences and identify seasonal churn patterns.

3. Ensemble Methods: Implement ensemble methods like stacking or blending to combine predictions from multiple models. Ensembles can often outperform individual models and provide more robust churn predictions.

## 4. Explainable AI: Utilize explainable AI techniques to improve model interpretability. Understanding the factors that contribute to churn prediction can help businesses take targeted actions to retain at-risk customers.

5. Customer Segmentation: Divide customers into meaningful segments based on similar characteristics or behaviours. Train separate models for each segment, as different customer groups may have distinct churn patterns.

## 6. Online Learning: Implement online learning techniques to continuously update the model with real-time data. This ensures the model remains current and adapts quickly to changes in customer behaviour.

7. Incorporate NLP and Sentiment Analysis: If available, include customer feedback or comments data to extract sentiment features. Sentiment analysis can provide valuable insights into customer satisfaction and identify potential churn triggers.

8. Contextual Information: Include external data sources, such as economic indicators or industry-specific trends, to provide contextual information that might influence customer churn.

## **9. Imbalanced Data Handling:** If the churn dataset is imbalanced (i.e., there are significantly more non-churners than churners), employ techniques like oversampling, under sampling, or using SMOTE to balance the data and prevent the model from being biased towards the majority class.

**10. Reinforcement Learning:** Consider using reinforcement learning techniques to optimize retention strategies. By learning from customer interactions and feedback, the model can make personalized recommendations to reduce churn.

**11. Customer Lifetime Value (CLV) Modelling:** Integrate customer lifetime value modelling with churn prediction to focus retention efforts on high-value customers who might be at risk of churning.

**12. A/B Testing:** Conduct A/B testing to validate the effectiveness of churn prevention strategies. This helps in fine-tuning the model and identifying the most successful interventions.

**13. User-Friendly Interface:** Develop a user-friendly dashboard or interface for business stakeholders to interact with the churn prediction model and access churn risk scores for individual customers.

**14. Deployment in Cloud:** Deploy the churn prediction model in a cloud-based environment to ensure scalability, flexibility, and easy access across multiple platforms.

**15. Feedback Loop:** Implement a feedback loop where the model's predictions are continuously monitored, and the model is retrained periodically to adapt to changing customer behaviour and business dynamics.

By incorporating these future enhancements, businesses can improve the accuracy and effectiveness of their churn prediction models, leading to better customer retention strategies and enhanced business performance.

17. Acknowledgment:

I would like to acknowledge the creators of the dataset used in this project, as well as the open-source machine learning libraries that facilitated the implementation of various algorithms and specially thanks to LearnWik that helps me to gain some experience while doing this project.

18. References:

## [***https://scikit-learn.org/stable/***](https://scikit-learn.org/stable/)

## [***https://www.w3schools.com/python/pandas/default.asp***](https://www.w3schools.com/python/pandas/default.asp)

[***https://www.w3schools.com/python/matplotlib\_intro.asp***](https://www.w3schools.com/python/matplotlib_intro.asp)

***https://www.w3schools.com/python/numpy/default.asp***