Project 5: CUSTOMER CHURN PREDICTION

ALGORITHM OR STEPS TO PREDICT CUSTOMER CHURN AND IDENTIFY FACTORS INFLUENCING CUSTOMER RETENTION USING RANDOM FOREST ALGORITHM:

Step 1: Data Collection and Preprocessing

Collect and gather relevant data, which may include customer demographics, transaction history, customer service interactions, and any other relevant information.

Preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features.

Step 2: Data Exploration

Perform exploratory data analysis (EDA) to understand the characteristics of your dataset.

Visualize the data using plots and summary statistics to identify patterns and potential factors influencing customer churn.

Step 3: Feature Selection

Identify which features (independent variables) are most likely to influence customer churn.

Feature selection techniques like correlation analysis, feature importance from Random Forest, or domain knowledge can help in this step.

Step 4: Split Data into Training and Testing Sets

Split your dataset into two parts: a training set and a testing set. A common split ratio is 70-30 or 80 20.

Step 5: Build the Random Forest Model

Train a Random Forest classifier on the training data.

Configure hyperparameters such as the number of trees, maximum depth of trees, and minimum samples per leaf. You can use techniques like cross-validation to find optimal hyperparameters.

Step 6: Model Evaluation

Evaluate the model's performance on the testing dataset using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and ROC AUC.

Analyze the confusion matrix to understand false positives and false negatives.

Step 7: Interpret Feature Importance

Use the feature importance scores from the Random Forest model to identify the factors that contribute most to customer churn.

Step 8: Fine-Tuning and Model Improvement

Experiment with different algorithms and hyperparameters to improve model performance

Consider techniques like oversampling or undersampling if dealing with imbalanced classes.

Step 9: Deployment and Monitoring

Deploy the trained Random Forest model into your production environment for real-time predictions.

Continuously monitor the model's performance and retrain it as needed with new data.

Step 10: Root Cause Analysis

Conduct further analysis to understand the root causes of customer churn based on the insights from the model. This may involve qualitative research and deeper investigation into specific customer segments.

Step 11: Implement Retention Strategies

Based on the identified factors influencing churn, develop and implement customer retention strategies. These strategies might include targeted marketing campaigns, improved customer service, or loyalty programs.

HARDWARE REQUIREMENTS FOR CUSTOMER CHURN PREDICTION:

CPU:

For small to medium-sized datasets and simple models, a standard multicore CPU can be sufficient.

For larger datasets and more complex models, a high-performance multicore CPU or a CPU cluster may be necessary to speed up training.

Memory (RAM):

The amount of RAM required depends on the size of your dataset and the complexity of your machine learning model. In general, having more RAM allows you to work with larger datasets and more extensive feature engineering.

16 GB or more of RAM is recommended for most machine learning tasks, but larger datasets may require 32 GB or even 64 GB.

GPU (Graphics Processing Unit):

GPUs can significantly accelerate training of deep learning models, such as neural networks. If you plan to work with deep learning models, having a GPU is highly beneficial.

NVIDIA GPUs are widely used in the machine learning community, and frameworks like TensorFlow and PyTorch have GPU support.

Storage:

You'll need sufficient storage space to store your dataset, model checkpoints, and any intermediate results. The exact storage requirements depend on the size of your data.

For very large datasets, consider using SSDs (Solid State Drives) or distributed file systems for faster data access.

Distributed Computing (Optional):

If you are dealing with extremely large datasets or complex distributed machine learning tasks, you may need a cluster of machines with distributed computing frameworks like Apache Spark or Hadoop.

Cloud Services:

Many cloud service providers (e.g., AWS, Google Cloud, Azure) offer machine learning services and GPU instances that can be easily provisioned as needed. This can be a cost-effective way to access high-performance hardware.

Network Connectivity:

Ensure you have a stable and fast internet connection, especially if you plan to use cloud-based services or download large datasets.

Cooling and Power:

Machine learning tasks can be computationally intensive and generate heat. Adequate cooling is essential to prevent overheating.

Ensure a stable power supply, as machine learning tasks can be time-consuming, and unexpected power interruptions can disrupt your work.

SOFTWARE REQUIREMENTS FOR CUSTOMER CHURN PREDICTION:

Python: Python is the primary programming language for data science and machine learning tasks.

Integrated Development Environment (IDE):

Jupyter Notebook or another Python IDE for coding and experimentation.

Machine Learning Libraries:

Scikit-learn: Scikit-learn is a widely-used Python library that includes the Random Forest algorithm and tools for building, training, and evaluating machine learning models.

Pandas: For data manipulation and analysis.

NumPy: For numerical operations on data arrays.

Data Visualization Libraries:Matplotlib and Seaborn

Data Preprocessing Tools:

Scikit-learn's preprocessing functions can be used for feature scaling, encoding categorical variables, and handling missing data.

Version Control:

Git and a platform like GitHub or GitLab for version control and collaboration with a team.

Database/Storage:

If your data is stored in a relational database, you may need database connectors like SQLAlchemy.

For big data scenarios, tools like Apache Spark or Hadoop can be helpful for data preprocessing.

Virtual Environments:

Use virtual environments (e.g., conda or virtualenv) to manage dependencies and isolate project environments.

Documentation and Collaboration Tools:

Tools like Markdown for documentation and platforms like Slack or Microsoft Teams for collaboration and communication with team members.

Random Forest-Specific Documentation:

Familiarize yourself with the Random Forest algorithm and its parameters. Scikit-learn's documentation is a great resource for this.

Model Evaluation and Metrics:

Scikit-learn provides various metrics for model evaluation, including those relevant to Random Forest models.

Hardware specifications:

❖ Operating system: Windows ,Linux,MacOS

❖ RAM : 8 GB to 16 GB

❖ Hard disc or SSD: at least 256 GB

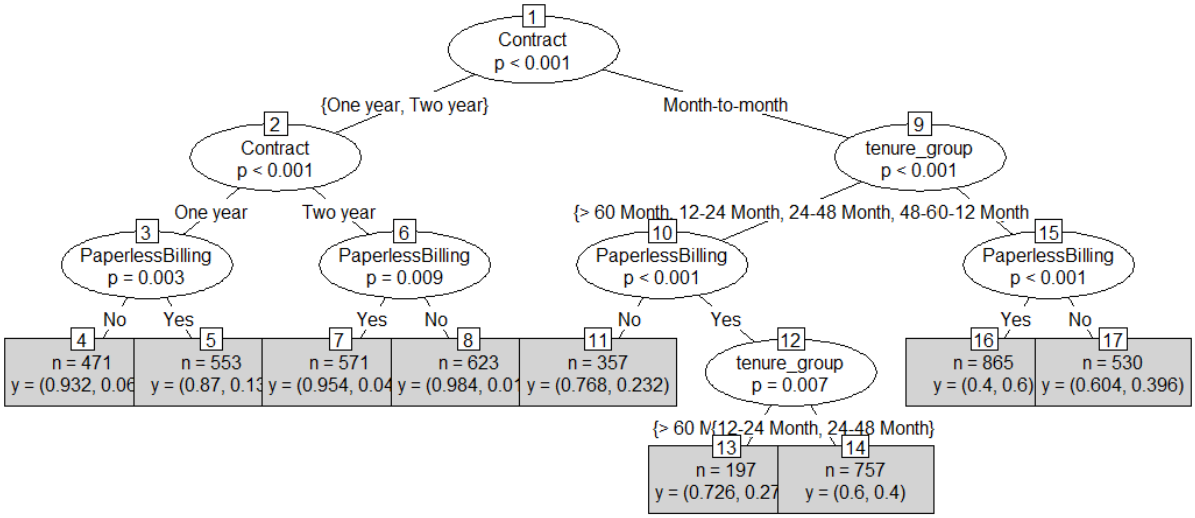
Software specifications:

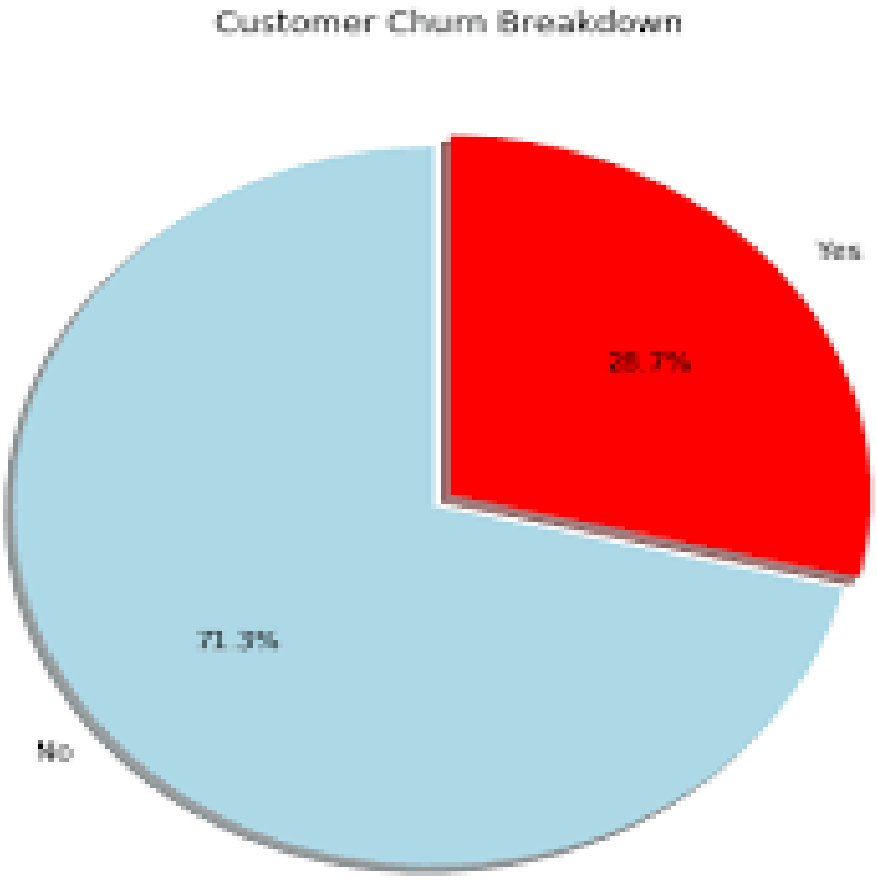
❖ Processor : Intel 3rd generation or high or Ryzen with 8 GB Ram

❖ Software’s : Python 3.6 or high version

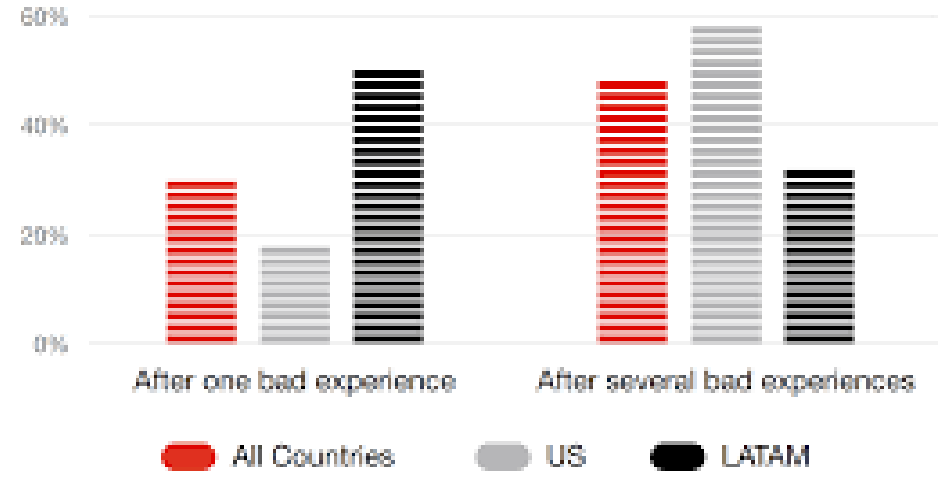
❖ IDE: Jupiter Notebook

FLOWCHART:

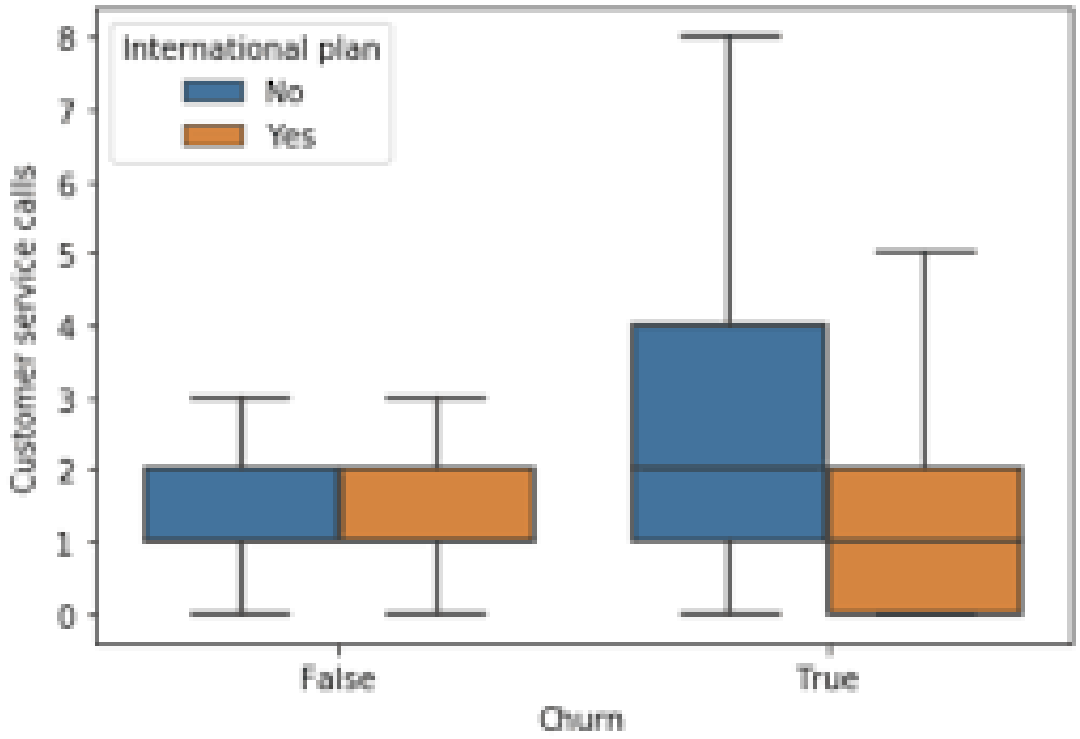

1
Contract
p < 0.001
{One year, Two year}
Month-to-month
2
Contract
p < 0.001
9
tenure_group
p < 0.001
One year
Two year
{> 60 Month, 12-24 Month, 24-48 Month, 48-60-12 Month
3
PaperlessBilling
p = 0.003
6
PaperlessBilling
p = 0.009
10
PaperlessBilling
p < 0.001
15
PaperlessBilling
p < 0.001
No
Yes
Yes
No
No
Yes
Yes
No
4
n = 471
y = (0.932, 0.06
5
n = 553
y = (0.87, 0.13
7
n = 571
y = (0.954, 0.04
8
n = 623
y = (0.984, 0.01
11
n = 357
y = (0.768, 0.232)
12
tenure_group
p = 0.007
16
n = 865
y = (0.4, 0.6)
17
n = 530
y = (0.604, 0.396)
{> 60 M{12-24 Month, 24-48 Month}
13
n = 197
y = (0.726, 0.27
14
n = 757
y = (0.6, 0.4)

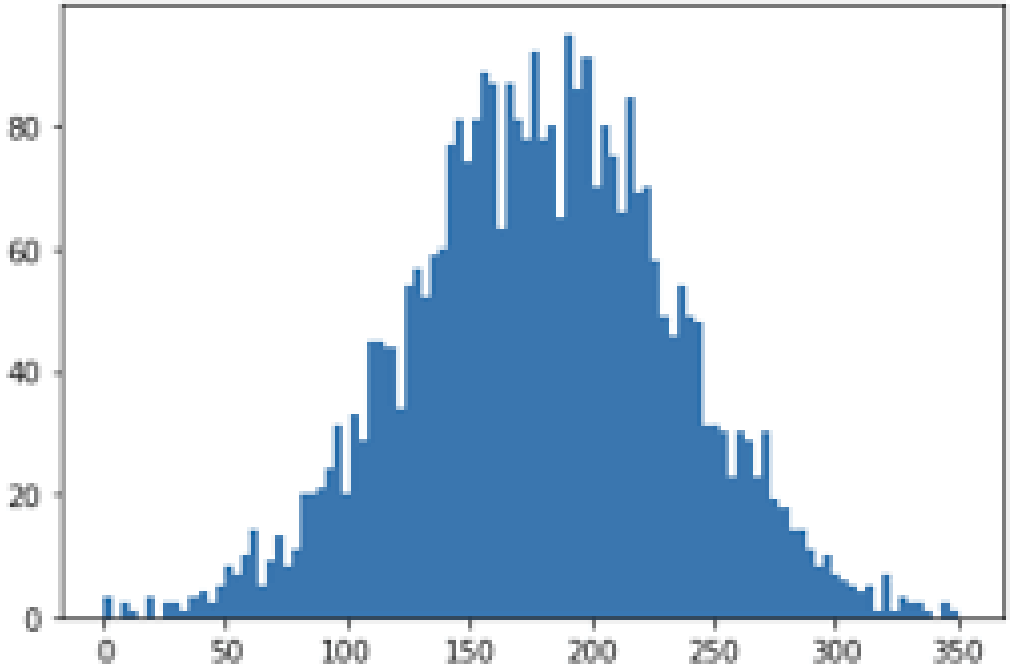


EXPECTED OUTPUT:



Q: At what point would you stop interacting with a company that you love shopping at or using?
Source: PwC Future of Customer Experience Survey 2017/18


International plan
No
Yes
Customer service calls
0
1
2
3
4
5
6
7
8
False
True
Churn


0
20
40
60
80
0
50
100
150
200
250
300
350