josephndex /
SyriaTel-Telecommunications-Customer-Churn-Prediction-Analysis

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

0 stars 0 forks 1 watching Branches Activity
Tags

Public repository

1 Branch 0 Tags Go to file t Go to file Add file Code

josephndex Add files via upload	f56cea6 · 3 minutes ago	
README.MD	Add files via upload	3 minutes ago
SyriaTel Telecommunications....	Add files via upload	3 minutes ago
bigml_59c28831336c6604c8...	Add files via upload	3 minutes ago
jupyter.ipynb	Add files via upload	3 minutes ago

README

SyriaTel Telecommunications company churn predictions

A study by Frederick Reichheld of Bain & Company,found that acquiring a new customer can be anywhere from five to 25 times more expensive than retaining an existing one. Additionally, Harvard Business Review has highlighted similar findings, emphasizing the value of keeping existing customers happy

1. INTRODUCTION :BUSSNESS PROBLEM

What is Customer Churn?

Customer churn occurs when customers or subscribers stop doing business with a company or service.

- In the telecom industry, customers have a wide range of service providers to choose from and often switch between them. This competitive market sees an annual churn rate of 15-25 percent.

- Individualized CUSTOMER RETENTION can be challenging because most companies have a large number of customers and cannot afford to dedicate significant time to each one. The costs would be too high, outweighing the additional revenue. However, if a company could predict which customers are likely to leave ahead of time, it could focus its retention efforts on these "high-risk" clients. The ultimate goal is to expand its coverage area and boost customer loyalty. The key to success in this market lies in the customer itself.

- CUSTOMER CHURN is a critical metric because retaining existing customers is much LESS EXPENSIVE than acquiring new ones.To REDUCE customer churn, telecom COMPANIES NEED TO PREDICT WICH CUSTOMERS HAVE A HIGHER CHANCE OF LEAVING. To detect early signs of potential churn, companies must develop a holistic view of their customers and their interactions across various channels, including store visits, product purchase histories, customer service calls, web-based transactions, and social media interactions.

- By addressing churn, these businesses can not only maintain their market position but also grow and thrive. The more customers they have in their network, the lower the cost of initiation and the higher the profit. As a result, reducing client attrition and implementing effective retention strategies are key focuses for a COMPANYS'S SUCESS.

2.OBJECTIVES

- 1.Churn Rate: Determine the percentage of customers who churn versus those who remain with active services.
- 2.Geographical Insights: Identify the states with the highest number of customers and those with the most churned customers.
- 3.Service Plan Impact: Assess the impact of service plans (international and voice mail) on customer churn.
- 4.Usage Patterns: Examine the correlation between usage patterns (day minutes, day charges, international charges) and customer churn
- 5.Customer Service Interaction: Analyze the relationship between customer service interactions and churn.
- 6.Feature Analysis: Investigate the correlation between features and customer churn.
- 7.Model creation: create models to predict customers who are more likely to churn

DATA UNDERSTANDING AND VISUALIZATION

- 1.Churn Imbalance: There is a significant imbalance between customers who churn and those who do not, with only 14% of customers churning.
- 2.Top States with Most Customers: The states with the highest number of customers are:

WV: West Virginia

MN: Minnesota

NY: New York

AL: Alabama

WA: Washington

- 2.States with Most Churned Customers: The states with the most churned customers are:

NJ: New Jersey

TX: Texas

MD: Maryland

MI: Michigan

NY: New York

- 3.Account Length: There is no correlation between account length and customer churn.
- 4.Area Codes: Area codes do not influence customer churn.
- 5.International Plan: Customers with an international plan are more likely to churn, indicating dissatisfaction with the service.
- 6.International Plan Adoption: Most customers do not have an international plan.
- 7.Voice Mail Plan: Customers with a voice mail plan are less likely to churn, suggesting satisfaction with the service.
- 8.Voice Mail Plan Adoption: Most customers do not have a voice mail plan.
- 9.Voice Mail Messages: Customers with fewer voice mail messages are more likely to churn, indicating satisfaction with the service when they use it more.
- 10.Day Minutes: Customers with high day minutes are more likely to churn, showing dissatisfaction with the service.
- 11.Total Day Calls: There is no correlation between total day calls and customer churn.
- 12.Daily Charges: Customers with high daily charges are more likely to churn, indicating dissatisfaction with the charges.
- 13.Evening Calls: Evening calls have no correlation to customer churn.
- 14.Night Minutes: There is no correlation between night minutes and customer churn.
- 15.Night Calls: There is no correlation between night calls and customer churn.
- 16.Night Charges: There is no correlation between night charges and customer churn.
- 17.International Minutes: International minutes have no correlation to customer churn.
- 18 International Calls: Customers with fewer international calls are more likely to churn.
- 19 International Charges: Customers with high international charges are more likely to churn, indicating dissatisfaction with the charges.
- 20.Customer Service: Customers who contact customer service are more likely to churn, suggesting dissatisfaction with customer service.

By focusing on these insights, you can better understand the factors influencing customer churn and develop targeted strategies to improve customer retention and satisfaction.

modeling

- I noticed that diffrent colums had diffrent scales and distributions

- we also have categorical values that need to be encoded
- ive also seen high cardinality categorical features
- solution> we are going to use powertransfomer to remove the skewness in our numerical columns and standard scaller to bring them to the same scale and one hot encoder to encode them and using frequency encoding for cardinality categorical features

MACHINE LEARNING MODEL AND EVALUATION

- we are now going to build our clasification models to predict wich customers are more likely to churn
- the main key focus is to reduce the number of false negatives while increasing the number of true positives

basic model

Here we are gona train the model with no tuning or any special changes to the data this will be our basic model

EVALUATION

here we are going to focus on 3 evaluation.

- our main focus is increasing the number of TRUE POSITIVES (customers more likely to churn) while reducing the number of FALSE NEGATIVES (customers who are likely to churn but the model cant identify htem).
- we saw that our data is imbalanced.
- so were gona focus on this key metrics for evaluation:
- 1.precision: we want to know how many of the predicted true positives are actualy true.
- 2.recall: we want to how many of the actual true positives are predicted true positives.
- 3.Number of true positives and negatives

Summary of Findings

Summary of Findings

1. Best Model: Gradient Boosting (0.85 ROC-AUC)

- Highest recall for Class 1 (71%) → Best at identifying minority class.
- High precision (90%) → Well-balanced model.
- Has 28 false negatives and 69 true positives

2. Random Forest (0.81 ROC-AUC)

- Better than Decision Tree and AdaBoost in overall performance.

- High precision for Class 1 (90%) but recall is lower (64%) → Can miss some minority cases.
- Has 34 false negatives and 63 true positives

3. Decision Tree (0.79 ROC-AUC)

- Performs well but less stable than Random Forest.
- Precision for Class 1 (65%) and a recall of (66%)
- Has 33 false negatives and 64 true positives

4. AdaBoost (0.64 ROC-AUC)

- Has a precision of (52%) for class 1
- Weaker recall for class 1 (34%) → Struggles to identify minority class.
- has 64 false negatives and 33 true positives

5. Logistic Regression (0.60 ROC-AUC)

- Very low recall for Class 1 (23%) → Poor at detecting minority class.
- Has 75 false negatives and 22 true positives
- Performs worst among all models in handling imbalanced data.

Key Takeaways

- Gradient Boosting is the best overall model (highest recall, and and has the least number of false positives).
- Random Forest is strong but weaker in recall for Class 1.
- Decision Tree works but overfits compared to ensembles.
- AdaBoost and Logistic Regression struggle with class imbalance.

Next Steps

- hyperparameter tunning
- Feature selection and engineering to improve recall.
- Consider SMOTE or class weighting to handle imbalance better.

MODEL USING HIGHLY SORILATED COLUMNS AND CLASS BALANCING

- we are going to balance the class using smote
- we are going to use a threshold of 0.10 for feature selection
- we are going to to use precision, recall to evaluate

EVALUATION

- the best are Random Forest and Gadient Boosting.
- Random forest model has with a recall and precision of (74%) with 25 false negatives and 72 true possitives.
- Gradient boosting has the ssame recal and precision as random forest eith the same number of false negatives and 72 true possitives.
- It slightly outperformed Gradient Boosting, especially in precision for churn cases (0.74 vs. 0.70) while maintaining the same recall (0.74).
- The models can be further optimized through hyperparameter tuning and feature importance analysis.

NEXT STEPS

- were are now going to focous on our 2 best performing models
- random forest and gadient boosting
- perform feature importance analysis to understand which features are most important for predicting churn

MODEL USING FEATURE SELECTION

SAMMARY

- from feature selection we have seen that our best model is random forest
- it has a recal of (77%) for class 1 and it has the least false negative 22

. HYPER PARAMETER TUNNING ON RANDOM FOREST ,GRADIENT BOOSTING and LIGHTGBM

-here we are going to focus on our 3 best models

- Random Forest has a slightly higher recall for Class 1 (0.76 vs. 0.75), meaning it catches more positives but has slightly lower precision.
- Gradient Boosting has slightly better precision (0.74 vs. 0.70) but misses more positives (higher false negatives).
- LightGBM False negatives are slightly higher
- Random Forest is likely the best model so far, as it offers a balance between recall and precision ..

we can now confidently conclude that our best model to predict wich customer is morelikely to churn Random forest

with a class 1 recall of 76%

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- Jupyter Notebook 100.0%