Isaac-Ndirangu-Muturi-7 ≡ phase 3 Public

<> **Code** Issues Pull requests Actions Projects Wiki Security Insights ···

main ▾

Commits on May 22, 2023

Create notebook.pdf
Isaac-Ndirangu-Muturi-749 committed 5 minutes ago

Update index-checkpoint.ipynb
Isaac-Ndirangu-Muturi-749 committed 1 hour ago

Update index.ipynb
Isaac-Ndirangu-Muturi-749 committed 1 hour ago

Create presentation.pdf
Isaac-Ndirangu-Muturi-749 committed 2 hours ago

Update index.ipynb
Isaac-Ndirangu-Muturi-749 committed 2 hours ago

Update index-checkpoint.ipynb
Isaac-Ndirangu-Muturi-749 committed 2 hours ago

Commits on May 21, 2023

Update README.md
Isaac-Ndirangu-Muturi-749 committed 2 days ago

Create customer-churn.jpg
Isaac-Ndirangu-Muturi-749 committed 2 days ago

Delete customer churn.jpg
Isaac-Ndirangu-Muturi-749 committed 2 days ago

Update README.md
Isaac-Ndirangu-Muturi-749 committed 2 days ago

Create customer churn.jpg
Isaac-Ndirangu-Muturi-749 committed 2 days ago

Delete customer churn.jpeg
Isaac-Ndirangu-Muturi-749 committed 2 days ago

Update README.md
Isaac-Ndirangu-Muturi-749 committed 2 days ago

Create reduce-customer-churn.jpg
Isaac-Ndirangu-Muturi-749 committed 2 days ago

Create customer churn.jpeg

Isaac-Ndirangu-Muturi-749 committed 2 days ago

phase 3

Update README.md

Isaac-Ndirangu-Muturi-749 committed 2 days ago

Update index.ipynb

Isaac-Ndirangu-Muturi-749 committed 2 days ago

update ipynb

Isaac-Ndirangu-Muturi-749 committed 2 days ago

Update README.md

Isaac-Ndirangu-Muturi-749 committed 2 days ago

Create bigml_59c28831336c6604c800002a.csv

Isaac-Ndirangu-Muturi-749 committed 2 days ago

Commits on May 17, 2023

Initial commit

Isaac-Ndirangu-Muturi-749 committed 5 days ago

Newer Older

Isaac-Ndirangu-Muturi-7 ☰ phase 3 Public

this project was part of the curriculum at a data science bootcamp - moringa-phase-3-project-data-science

☆ **0** stars **0** forks

☆ Star Watch

Code Issues Pull requests Actions Projects Wiki Security Insights ···

main ▾ ···

Isaac-Ndirangu-Muturi-749 Create notebook.pdf ··· 8 minutes ago **21**

View code

README.md

SYRIATEL CUSTOMER CHURN ANALYSIS

This project was part of the curriculum at a data science bootcamp at moringa school, phase 3 project.

OVERVIEW

This project aims to analyze and predict customer churn in the telecommunications industry. The goal is to develop a machine learning model that can accurately predict whether a customer is likely to churn or not. By identifying potential churners, the telecommunications company can take proactive measures to retain customers and minimize revenue loss.

The analysis includes exploratory data analysis, feature engineering, model selection, and evaluation. Various machine learning algorithms, such as Random Forest, Decision Tree, K-Nearest Neighbors, Logistic Regression, and SVM, were trained and evaluated based on their accuracy, F1 score, recall, and precision.

Features

phase 3

- Exploratory data analysis to understand the dataset and identify patterns.
- Feature engineering to create relevant features for modeling.
- Training and evaluation of multiple machine learning algorithms.
- Cross-validation techniques to ensure robustness of the models.
- Evaluation metrics including accuracy, F1 score, recall, and precision.
- Feature importance analysis to identify the most influential factors for customer churn.
- Visualizations such as ROC curves, bar plots, and confusion matrices to present the results.

BUSINESS UNDERSTANDING

As a data scientist assigned to investigate customer churn for SyriaTel, my main objective is to analyze the available data and develop a predictive classifier that can accurately determine whether a customer is likely to terminate their relationship with the telecommunications company. By understanding the underlying patterns and reasons behind customer churn, our aim is to assist SyriaTel in reducing financial losses and implementing targeted retention strategies. Through comprehensive data analysis and modeling techniques, we can identify key factors influencing churn and provide actionable insights to the business. The stakeholder audience for this project consists of key decision-makers and stakeholders within the telecommunications industry, such as executives, marketing managers, and customer retention teams. These individuals are responsible for strategic planning, customer acquisition, and implementing measures to reduce customer churn.

To achieve this goal, I will begin by conducting a thorough examination of the dataset, encompassing customer demographics, usage patterns, billing information, and customer service interactions. This exploratory analysis will enable me to gain a deep understanding of the data, identifying potential features that have a significant impact on customer churn. By leveraging statistical techniques and visualization methods, I can uncover correlations and patterns that will serve as the foundation for the subsequent modeling phase.

Once the dataset has been carefully examined, I will preprocess the data to handle missing values, encode categorical variables, and normalize numerical features. This preprocessing step is crucial to ensure the dataset is suitable for modeling, as it minimizes bias and enhances the quality of the input data. Additionally, I will employ feature selection techniques to identify the most relevant variables or engineer new features that can provide valuable insights into customer churn. This process will involve assessing feature importance, conducting correlation analysis, and incorporating domain knowledge expertise to select the most informative set of features.

After feature selection and engineering, I will select an appropriate machine learning algorithm for the classification task. Depending on the nature of the data and the problem at hand, algorithms such as logistic regression, decision trees, random forests, support vector machines (SVM), or gradient boosting algorithms like XGBoost or LightGBM may be considered. The chosen algorithm will be trained on the preprocessed dataset, employing suitable training techniques such as cross-validation to ensure the model's robustness and generalization capabilities. By iteratively refining the model's parameters and evaluating its performance, we can develop a reliable classifier for predicting customer churn.

DATA UNDERSTANDING

phase 3

The dataset choice is crucial in addressing the specific needs and interests of the stakeholder audience. The selected dataset contains comprehensive information about customers, including demographic details, call patterns, and account information, which are essential factors in understanding customer behavior and predicting churn.

By utilizing this dataset, stakeholders can gain valuable insights into customer churn patterns and make informed decisions to enhance customer retention strategies. The dataset's diverse range of features allows for a more holistic understanding of the factors influencing churn, enabling stakeholders to identify key drivers and take appropriate actions to mitigate churn rates.

The data understanding phase involved exploring and analyzing the dataset to gain insights into its structure, features, and characteristics. Here is a summary of the key findings:

Dataset Overview:

The dataset contains information about customers in the telecommunications industry, including various features and a target variable indicating churn status. It consists of a significant number of records, providing a robust foundation for analysis and modeling.

Features:

The dataset contains a diverse range of features, including customer demographics (e.g., state, area code), call details (e.g., number of calls, call duration), and account information (e.g., international plan, voice mail plan). These features offer a comprehensive view of customer behavior and characteristics that can potentially impact churn.

Target Variable:

The target variable, "churn," indicates whether a customer has churned or not. The class distribution shows an imbalance, with a relatively lower proportion of churned customers compared to non-churned customers. Addressing the class imbalance may be necessary during the modeling phase to ensure reliable predictions.

Data Quality:

The dataset appears to be relatively clean, with no major issues such as missing values or significant outliers. However, further analysis and preprocessing may be required to handle categorical variables, standardize numeric features, and address any other specific requirements of the modeling algorithms.

Exploratory Data Analysis:

Initial data exploration revealed potential relationships between certain features and the churn status. Some features, such as total day charge and customer service calls, exhibited noticeable differences between churned and non-churned customers. Correlation analysis and visualization techniques were used to identify potential predictors of churn. In summary, the data understanding phase provided a comprehensive understanding of the dataset, its features, and the target variable. This knowledge serves as a foundation for subsequent data preparation, modeling, and analysis, helping to uncover meaningful patterns and insights related to customer churn in the telecommunications industry.

MODELLING

During the modeling stage, five machine learning models were employed to predict customer churn in the telecommunications industry. Here is a summary of each model:

1. Logistic Regression:

This model uses a logistic function to predict the probability of binary outcomes. It achieved a moderate accuracy and F1 score, indicating reasonable predictive performance. Logistic Regression provides interpretability by examining the coefficients associated with each feature.

2. Random Forest:

This ensemble learning algorithm combines multiple decision trees to make predictions. It demonstrated the highest accuracy and F1 score among all the models, indicating its strong predictive capabilities. The top features identified by the Random Forest model were "total day minutes," "total day charge," "total eve calls," "total eve charge," "total night charge," "total intl calls," "customer service calls," "area_code_is_415," "voice_mail_plan_is_yes," and "churn."

3. Decision Tree:

This model utilizes a tree-like structure to make decisions based on feature values. It achieved a relatively high accuracy and F1 score, but slightly lower than the Random Forest model. Decision trees provide interpretability, allowing us to understand the rules and conditions used for predictions.

4. K-Nearest Neighbors (KNN):

KNN classifies data points based on their proximity to neighboring data points. It achieved a lower accuracy and F1 score compared to the Random Forest and Decision Tree models. KNN's performance is heavily influenced by the choice of K (number of neighbors) and the distance metric used.

5. Support Vector Machine (SVM):

SVM separates data points by maximizing the margin between different classes. It achieved a relatively lower accuracy and F1 score compared to the other models. SVM's performance is sensitive to the choice of kernel function and regularization parameters.

Overall, the Random Forest model outperformed the other models in terms of accuracy and F1 score, making it the most effective model for predicting customer churn in the telecommunications industry. However, the other models also provided valuable insights and can be useful in certain scenarios based on their unique characteristics.

EVALUATION

Based on the analysis using the Sequential Forward Selection (SFS) with a Random Forest model, the top 10 features that have the most significant impact on customer churn are as follows:

- Total day charge: This feature has the highest importance, indicating that the charges incurred during daytime usage play a crucial role in determining customer churn.
- Total eve charge: The charges for evening usage also have a significant impact on customer churn.

- Customer service calls: The number of customer service calls made by customers is a strong indicator of potential churn.
- Total night charge: Charges related to nighttime usage contribute to the likelihood of churn.
- Total day calls: The number of calls made during the day affects customer churn.
- State: The geographical location or state of the customer can influence churn behavior.
- Number vmail messages: The presence or absence of a voicemail plan and the number of voicemail messages impact churn.
- Voice_mail_plan_is_yes: Whether a customer has a voicemail plan or not affects churn behavior.
- Area_code_is_415: The area code "415" has a slight impact on churn.
- Area_code_is_408: The area code "408" also has a minor effect on churn.

Results

The analysis revealed that the Random Forest model achieved the highest accuracy and F1 score among the evaluated models. It showed a strong ability to predict customer churn, with an accuracy of 96.503% and an F1 score of 0.77477. The top features identified as influential in predicting churn were total day charge, total eve charge, customer service calls, total night charge, total day calls, state, number of voicemail messages, voice mail plan (yes/no), area code 415, and area code 408.

RECOMMENDATIONS

Based on the findings, SyriaTel can take the following actions to reduce customer churn and minimize revenue loss:

- Focus on managing and reducing the charges incurred by customers during daytime and evening usage. Analyze the pricing structure and consider offering competitive plans to retain customers.

phase 3

- Pay close attention to the number of customer service calls. Identify the root causes behind frequent calls and take proactive measures to address customer issues promptly.
- Monitor and optimize the nighttime charges to ensure they are aligned with customer expectations and market standards.
- Develop strategies to improve customer satisfaction and engagement during daytime usage, as indicated by the number of day calls.
- Understand the specific factors related to each state and tailor marketing efforts and customer retention initiatives accordingly.
- Evaluate the effectiveness of voicemail plans in retaining customers. Consider enhancing the features and benefits of these plans to increase customer loyalty.
- Analyze the customer churn patterns associated with different area codes. Identify any specific issues or challenges faced by customers in those areas and develop targeted retention strategies.
- Tailor marketing strategies: Utilize the information gained from the analysis to develop targeted marketing campaigns. By understanding the factors that contribute to customer churn, SyriaTel can tailor its marketing efforts to address customer needs and preferences more effectively.
- Monitor customer satisfaction: Regularly assess customer satisfaction levels through surveys, feedback mechanisms, and customer interactions. Identify and address potential pain points or areas where customers might be dissatisfied to proactively prevent churn.
- Leverage predictive models: Implement the tuned Random Forest model to predict customer churn in real-time. Continuously update and refine the model based on new data to improve its accuracy and effectiveness.

By implementing these recommendations, SyriaTel can reduce customer churn and mitigate the financial impact caused by customers who do not stay with the company for an extended period.

CONCLUSION

In this analysis, we explored several machine learning models to predict customer churn in the telecommunications industry. We compared the performance of various models, including Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors, and Support Vector Machine (SVM), after tuning their hyperparameters. The evaluation metrics used to assess the models were accuracy, F1 score, recall, and precision.

Among the models examined, the Random Forest model (tuned) outperformed other models in accurately predicting customer churn.

Furthermore, by applying the Sequential Forward Selection (SFS) technique to the Random Forest model, we identified the top 10 features that significantly contribute to predicting customer churn.

License

phase 3

Acknowledgments

The dataset used in this project is provided by SyriaTel. Special thanks to the open-source community for their valuable contributions to the libraries and tools used in this project.

Contact

For any questions or inquiries, please contact:
👉 Twitter: https://twitter.com/NdiranguMuturi1?t=xXF2OKsqOUeb5J_4yysFKg&s=09
👉 LinkedIn: https://www.linkedin.com/in/isaac-muturi-3b6b2b237
👉 GitHub: https://github.com/Isaac-Ndirangu-Muturi-749

Feel free to contribute to this project by submitting pull requests or opening issues. Your feedback is highly appreciated.

Happy analyzing and predicting customer churn!

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● **Jupyter Notebook** 100.0%