Customer Churn Prediction Project

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

EOkeyo Presentation and notebook conversion ded2ece · 3 minutes ago

Churn.csv	Project commit	43 minutes ago
Presentation.pptx	Presentation and notebook conversi...	3 minutes ago
README.md	Presentation and notebook conversi...	3 minutes ago
notebook.pdf	Presentation and notebook conversi...	3 minutes ago
project.ipynb	Project commit	43 minutes ago

README

Customer Churn Prediction Project

(Machine Learning Project (Phase 3))

Project Overview

This project focuses on predicting customer churn using machine learning models. Predicting churn is a critical business objective because retaining existing customers is far more cost-effective than acquiring new ones. By identifying customers at risk of leaving, the company can deploy proactive retention strategies such as personalized discounts, loyalty programs, or targeted engagement campaigns to preserve revenue and maintain market share. This analysis evaluates four models:

- Logistic Regression (baseline linear model)
- Decision Tree Classifier (rule-based model)
- Random Forest (ensemble of decision trees)
- Gradient Boosting (boosted ensemble model) The goal is to determine which model best identifies customers likely to churn while balancing interpretability and predictive performance.

Business and Data Understanding

Stakeholder Audience

The main stakeholders are the Customer Success and Retention Teams. They require a system that flags customers at risk of churn before they leave, enabling proactive interventions rather than reactive responses. This approach helps reduce revenue loss, improves customer lifetime value, and strengthens long-term relationships.

Dataset Choice

The project uses the churn.csv dataset, which contains 3,333 records describing customer behavior.

Features: -

- account length
- international plan
- voice mail plan
- number vmail messages
- total day minutes
- total day charge
- total eve minutes
- total eve charge
- total night minutes
- total night charge
- total intl minutes
- total intl charge
- customer service calls

Target Variable (churn):

- Class 0: Customer retained
- Class 1: Customer churned

Challenge:

- The dataset is imbalanced, with fewer churned customers than retained ones. This makes it harder for models to identify the minority class, so metrics beyond accuracy such as Recall, Precision, and F1-score are critical for evaluation.

Modeling Approach

Four classification models were trained and evaluated:

1.Logistics Regression:

- Linear model used as a baseline, providing interpretability by showing how each feature influences the probability of churn. ***2.Decision Tree Classifier:***
- Non-linear, rule-based model that can capture complex interactions between features. ***3.Random Forest:***
- An ensemble of decision trees improving predictive stability and reducing overfitting. ***4.Gradient Boosting:***
- A boosted ensemble model that iteratively improves weak learners, typically yielding the highest predictive accuracy.
- All models were trained on the same processed dataset to ensure fair comparison.

Evaluation Metrics

- Models were evaluated using Accuracy, Precision, Recall, F1-Score, and AUC.
- ***Accuracy:*** Overall percentage of correct predictions.
- ***Precision:*** Percentage of predicted churners who actually churned (reliability of alerts).
- ***Recall:*** Percentage of actual churners correctly identified; crucial for reducing revenue loss.
- ***F1-Score:*** Harmonic mean of Precision and Recall; balances false positives and false negatives.
- ***AUC (Area Under ROC Curve):*** Measures overall ability to distinguish churners from non-churners.

Model Performance

Confusion Matrix

- True positive = 43
- True negative = 1383
- False positive = 42
- False negative = 199

Accuracy

Model Accuracy Logistic Regression 85.5% Decision Tree 88.8% Random Forest 93.2% Gradient Boosting 93.2%

Interpretation:

- Random Forest and Gradient Boosting have the highest overall accuracy. However, due to class imbalance, accuracy alone does not fully reflect the ability to detect churners.

Precision, Recall & F1-Score

Model Precision Recall F1-Score Logistic Regression 0.506 0.178 0.263 Decision Tree 0.618 0.595 0.606 Random Forest 0.927 0.579 0.712 Gradient Boosting 0.861 0.616 0.718

Interpretation: • Logistic Regression: Misses most churners (Recall = 17.8%), making it unsuitable for proactive retention despite decent accuracy. • Decision Tree: Balanced performance (Recall = 59.5%, F1 = 60.6%), interpretable rules for business use. • Random Forest: Extremely precise (92.7%) and strong F1-score (71.2%), but moderate recall means some churners are still missed. • Gradient Boosting: Best overall balance (Recall = 61.6%, F1 = 71.8%), high AUC (0.913), and strong predictive performance. **AUC Score** Model AUC Logistic Regression 0.830 Decision Tree 0.766 Random Forest 0.906 Gradient Boosting 0.913

Interpretation:

- Gradient Boosting demonstrates the strongest ability to separate churners from non-churners, followed closely by Random Forest. Decision Tree and Logistic Regression are less effective at distinguishing classes.

Limitations

1. Class Imbalance: Fewer churners than retained customers may cause models to under-detect minority cases. Some churners could still be missed despite using ensemble methods.
2. Threshold Dependence: Metrics are based on a standard probability threshold (0.5). Adjusting thresholds could improve recall at the expense of precision.
3. Feature Limitations: Dataset contains limited features. External factors such as competitor pricing, promotions, or customer sentiment were not included and may impact churn behavior.
4. Overfitting Risk: Complex models like Random Forest and Gradient Boosting may overfit to training data if not monitored carefully, reducing generalization on new customer data.
5. Interpretability: Ensemble models (Random Forest, Gradient Boosting) provide less transparency compared to Decision Trees or Logistic Regression, which may complicate explaining decisions to non-technical stakeholders.

Conclusion

Key Insights • Logistic Regression: Poor recall, not suitable for churn prevention. • Decision Tree: Interpretable, moderate performance; good for explaining patterns to business teams. • Random Forest: Highly precise; excellent for targeted interventions. • Gradient Boosting: Best balance of recall, precision, F1-score, and AUC; most effective for proactive retention.

Recommendations

1. Deploy Gradient Boosting for production due to superior overall performance.
2. Use predicted probabilities to rank high-risk customers for targeted retention campaigns.
3. Integrate predictions into CRM systems for real-time alerts to account managers.
4. Adjust thresholds to increase recall if business priorities favor identifying more churners.
5. Future Work: Explore additional features, resampling methods, or alternative ensemble techniques to further improve detection of minority churn cases.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%