Phase 3 Project

Flatiron Online Data Science Course

Summary

This is the third end-of-module project in the Flatiron Online Data Science Course

The project aims at testing and reinforcing the fundamental **machine learning theory and algorithms** taught in phase 3 of the Flatiron School curriculum

Applying the tools of **logistic regression** enables **determining the likelihood of a given data point being associated with one of two categories**. In combination with regularization techniques and machine learning algorithms, it can be used to solve various different classification problems.

Outline

- Business Problem
- Data & Methodology
- Results
- Conclusions & Actionable Insights
- Further studies

Business Problem

The telecom business of SyriaTel is interested in reliable ways to forcast customer churn and hence either focus on its remaining customers or find ways to prolong customer lonjevity.

Forecasting whether or not a customer is likely to churn is a **binary classification** problem requiring models to predict potential customer churn as accurately as possible.

Data & Methodology

- The SyriaTel dataset consists of 3,333 entries with information on its US customer base
- After careful exploration and cleaning of the data, different variables were analyzed in order to reveal their predictive capabilities for customer churn
- Ultimately, the data was used to find the predictive model with the highest accuracy

Results

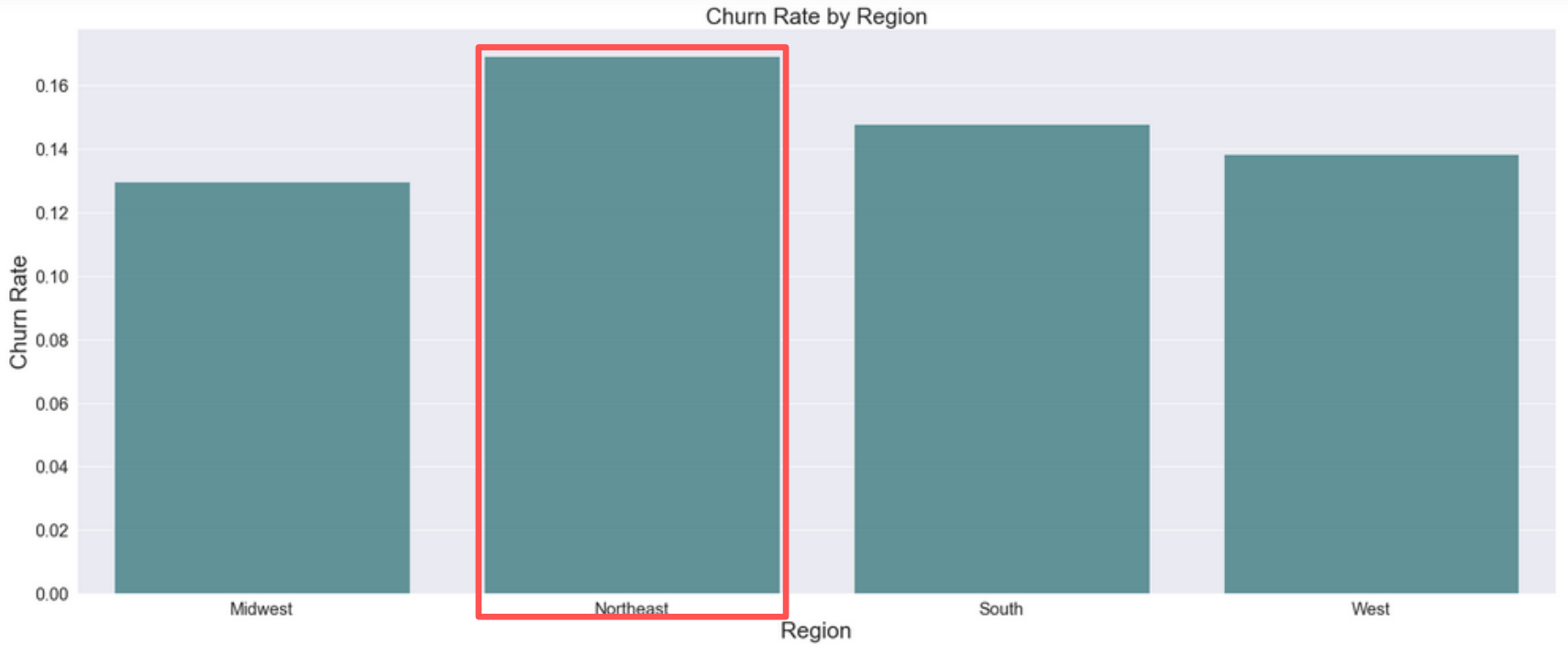
The following key features were explored throughout the project:

1. Region
2. Total Charge
3. Customer Service Calls
4. International Plan

After testing several models with different algorithms and comparing their predictive abilities, the model with the highest accurcacy has been identified to predict customer churn.

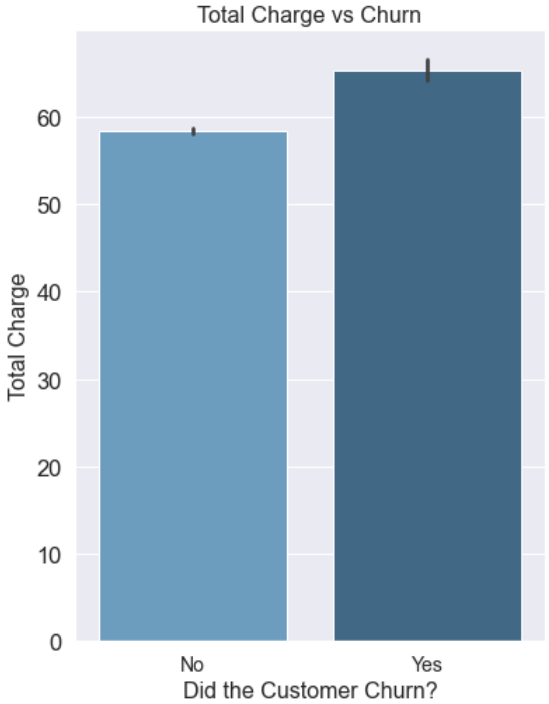
1. Region

- Is customer churn dependent on the location of the customer?



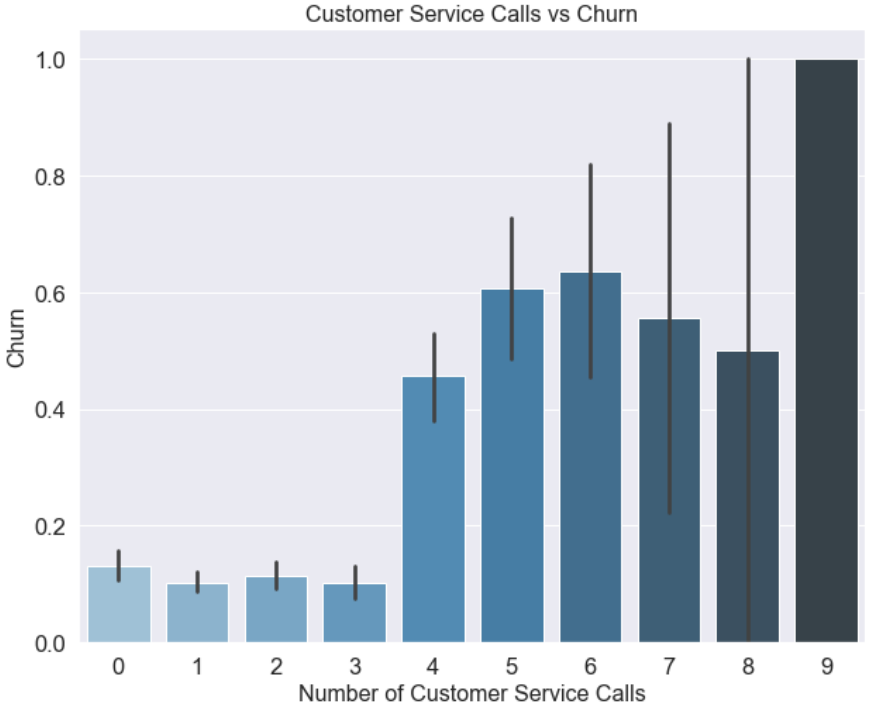
2. Total Charge

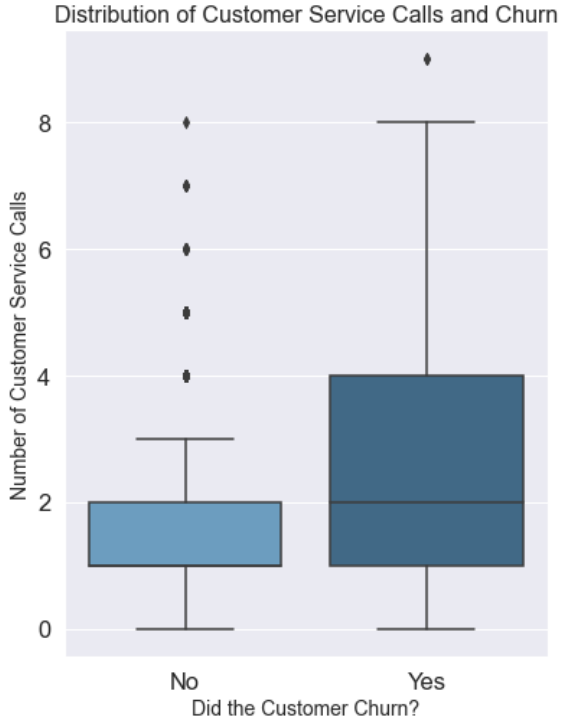
- How does total charge of phone services affect customer churn?



3. Customer Service Calls

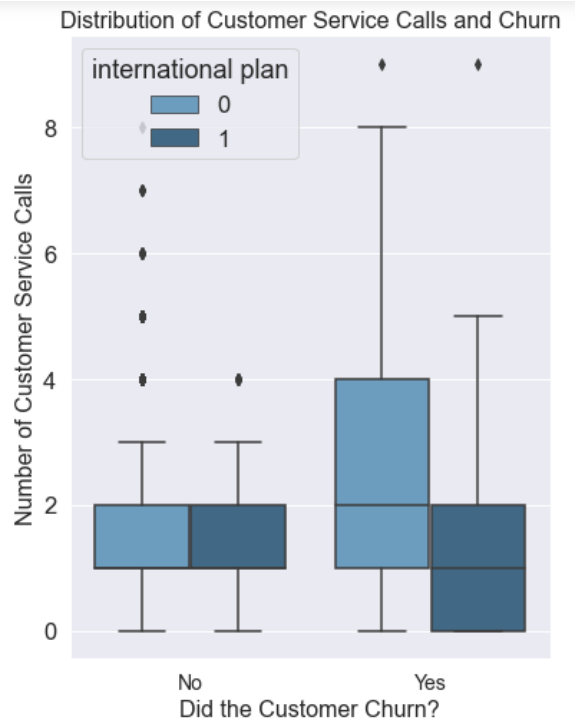
- What is the effect of customer service calls on customer churn?





4. International Plan

- Does the availability of an international plan affect customer churn?



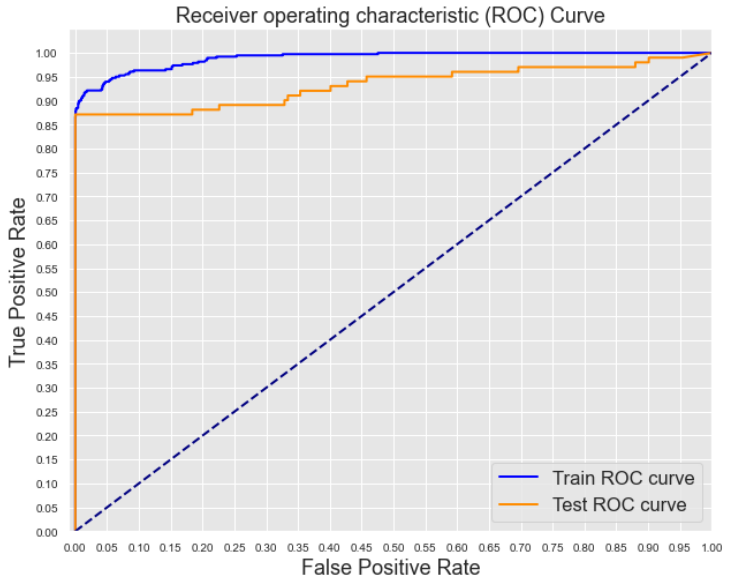
Predictive models

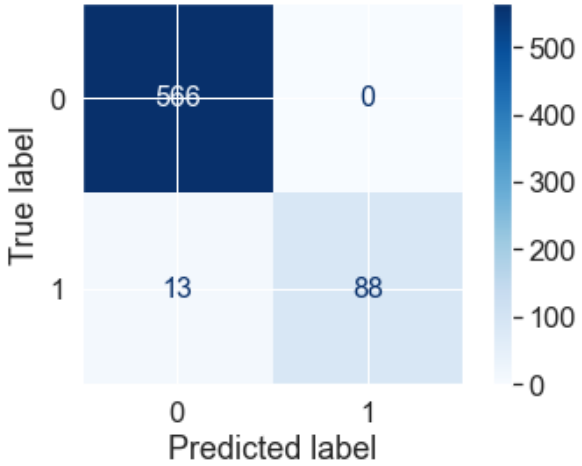
Model type	Description	Test Score	Train Score	Precision	Recall	F1-Score
Random Forest	Bagging	98.1%	98.0%	100%	87.1%	93.1%
XGBoost	Base	98.1%	98.3%	100%	87.1%	93.1%
Gradient Boost	Base	97.9%	98.3%	98.9%	87.1%	92.6%
Decision Tree	Tuning/Pruning	96.7%	97.7%	90.7%	87.1%	88.9%
Decision Tree	Base	95.1%	100%	80.9%	88.1%	84.4%
Random Forest	SMOTE	94.6%	94.9%	100%	63.4%	77.6%
Random Forest	Base	94.4%	95.5%	85.5%	70.3%	77.2%
Adaboost	Base	92.3%	92.6%	83.8%	61.4%	70.9%
KNN	Base	89.8%	92.2%	90.2%	36.6%	52.1%
KNN	Best k	88.1%	90.1%	93.5%	23.0%	36.9%
Logistic Regression	SMOTE	85.9%	89.7%	61.9%	25.7%	36.4%
KNN	SMOTE	78.2%	92.4%	38.9%	72.2%	50.6%
Logistic Regression	Base	77.9%	76.9%	38.8%	79.2%	52.1%

Best performing model

Random Forest with Bagging

- Accuracy: 98%
- F1-score: 93%






```
Testing Accuracy Score for the Bagged Decision Tree Classifier: 98.051%

Classification Matrix:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       566
           1       1.00      0.87      0.93       101

    accuracy                           0.98       667
   macro avg       0.99      0.94      0.96       667
weighted avg       0.98      0.98      0.98       667
```

Conclusions & Actionable Insights

1. Focus on **lower pricing** or cost benefits such as discounts

2. Ensure **higher quality customer service**, reaching a benign solution for the customer after three customer service calls in order to avoid an increased potential for churn

3. Focus analytical efforts on the **international plan** as it has been shown that it has a strong influence on whether or not a customer is retained.
    - Possible aspects to investigate are **availability and pricing**

Further Studies

1. **Gridsearch (Gradient Boost or XGBoost)**
   - Find the best combination or parameters

2. **Expand research and analysis by more variables**
   - E.g. Number of Voicemail messages or in-depth analysis of call lenghts

3. **Model refinements**
   - Analyze feature importance

Thank You!

Email: kontakt@oliverzimmer.eu
GitHub: @senseize
LinkedIn: linkedin.com/in/username/oliver-zimmer-cfa-8824881ab/