Business Analytics II first course assignment

Hektor Dahlberg (2201899)

October 26, 2024

Contents

1 Business Understanding

1.1 What is customer churn, and why is predicting it important for subscription-based businesses?

Customer churn is where customers stop/discontinue doing business with a company/service. Why it is important to predict it is because

1. Lost revenue less customers less revenue, and keeping existing customers is cheaper then trying to get new ones.

2. Resource managment Knowing why customers leave helps to improve the business to improve the parts why customers are leaving and increase the resources for making them better.

3. Marketing Knowing which customers are likely to leave helps in strategies to offer/improve customers satisfaction and as such lowers their churn rate.

1.2 What is the potential impact of reducing churn on a company's revenue and customer retention strategy?

1. Increased revenue In general more customers higher revenue, ofcourse if your strategy is to lower price so you can lower churn this maybe not true.

2. Steady customer relationship A low customer churn rate means your customers stay longer as customers, which lowers the likelihood that they will stop their subscription.

1.3 What could be the specific business objectives?

Finding the optimal churn rate to monthly/yearly charge where the churn rate is low and the income is high.

1.4 How can churn prediction affect marketing and customer service strategies?

By for example direct marketing or giving out better deals for customers who are highly likely to unsubscribe from your service.

2 Data Understanding

2.1 What data is available to predict churn

Customer demographics: Gender, Senior citizen, partner status, dependents
Subscription details: contract type, payment method, paperless
Usage behavior: Tenure, Monthly charges, Total charges, Internet plan, extra services

2.2 What are the key features or variables that might affect customer churn?

Seems to be that the usage behaviors are the key features with customer churn.

2.3 Are there any missing values or outliers in the dataset that need attention?

Yes there are, few Nan values for customers who have no total charges and the more problematic are the no phone service/ no internet plan.

2.4 What relationships can you observe between features and the target variable (churn)?

No one feature has a high correlation with churn there are a few features which have either an medium negative or positive correlation with churn.

2.5 Can you identify any early insights?

Customers who do not have a internet plan are less likely to churn same goes for 2 year contracts and a high total charge, while people who have a monthly plan are highly likely to churn.

3 Data Preparation

3.1 How will you handle missing data or outliers?

The 9 rows of NaN data i will just drop as it is such a small amount of data, for outliers I drop the also from the dataset but i could not find any.

3.2 Do you need to normalize/standardize any of the features?

Yes I decided to use MinMax scaler for the customer tenure, monthly charge and total charge, also because to total charge is highly skewed I just an logarithmic transformation to reduce the skewness, this also can explain why there are no outliers afterwards when i look for them as it also helps with handling of outliers.

3.3 What methods will you use to split the dataset into training and testing sets?

Just a normal train test split of 20/80.

3.4 Should any features be excluded based on relevance or redundancy?

I thought so, but after testing models with different feature sets(one with only features that have an feature importance higher than 0.04, only usage behaviour features, usage and subscription details and only customer demographics), the best results came when I included all features. It seems counterintuitive that the best resluts came from when i used all the dataset as it usually removing irrelevant features introduce noise/ overfitting, can be that my selection of features are sub optimal.

4 Model Building

4.1 What measure(s) will you use to evaluate the performance of the models?

accuracy as when i run the grid search i try to find the best parameters for this measure.

4.2 Create a baseline model as a logistic regression without any parameter optimization

Done.

4.3 Create various tree-based models, also experiment with parameter optimisation: (i) decision trees, (ii) bagging, and (iii) random forest classifiers.

Done

5 Evaluation

5.1 How did the models perform overall? What were the strengths and weaknesses of each model?

All of them had about the same result, no matter which features i used they got about the same test accuracy, with all the data the accuracy for the test run was between 0.75 to 0.79.

5.2 Which model performed best based on the performance evaluation metrics you chose, and why do you think it outperformed the others?

Logistic regression performed the best, why I really am not sure about. Maybe the underlying relationship in the data is approximately linear, less overfitted data as more complex models tend to overfit easily (tried to minimaize this but can be that just simply failed to pre process the data well for the other models).

5.3 Would the model(s) generalize well to new, unseen data?

Depends on the size and quality of the unseen data. As I suspect my other models then the logistic regression model are capturing some noise instead of the patterns so when the underlying relationships in the data are not overly complex I believe that the logistic regression model would work better with ”simple” data while more ”complex” data and bigger data sets an more complex model should work better.

5.4 What are the four most important predictors according to the best decision tree model?

Monthly charges, Total charges, Contract two year and Customer Tenure

6 Deployment

6.1 How would you implement this model in a real business environment?

Try integrating it into the existing customer relationship management system, and make sure that data pipelines are setup to ensure that we can have real-time predictions of churn. At the same time with increased data we need to continuously monitor model performance and update/change it if necessary.

6.2 What steps would the company need to take to integrate this model into their customer management system?

same answer as before.

6.3 How could the model be used for targeted interventions (e.g., offering discounts to customers likely to churn)?

Identify customers predicted to churn and design personalized campaigns, targeted discounts or special offers. Try to proactively engage with the customers who are at-risk to churn, also from this data we can see which methods of targeted intervention works best and can help us on our future decision when deciding on our strategy to keep customers from churning.

6.4 What potential risks or limitations could arise during deployment (e.g., overfitting, ethical concerns, data drift)?

Potential risks include overfitting, where the model fails to generalize to new data, and data drift, where changing customer behavior reduces prediction accuracy. Ethical concerns are not that high in my opinion as due to the non-sensitive nature of the data, but of course as users under GDPR have rights regarding their personal data, including the right to restrict processing, so users can request that their data is not used for training ML-models, this can be a problem if a large amount of users opt-out from sharing their data with us. which then can lead to worse model performance.