Nikita Goldovsky

Technical Exercise

What questions would you want to ask about how this data set was produced?

-Is everyone in the dataset up for renewal?

-Is there any intuition about how null values should be treated in the analysis?

-Does this dataset represent a specific sample of the overall audience? I.e. will the inference of the analysis apply to the entire audience?

Before you begin the analysis, what types of things would you expect to influence the likelihood of whether or not a subscriber renews?

-How active the user has been recently (# of visits per day, # of days per week with a visit).

-How long the user has been a paid subscriber.

-Whether the user is on autopay.

-Whether it is the user or his/her company that is paying.

After exploring the dataset, what do you notice that might influence how you approach the analysis and modeling?

The variables **Channel**, **AutoRenewOn**, **Company**, and **DaysSinceLastLogin** have NULL values that need to be accounted for.

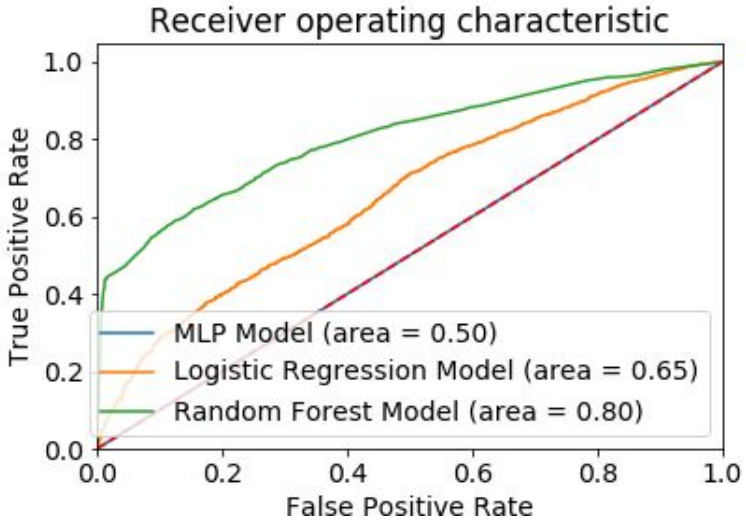
There is a mix of numerical and categorical data.

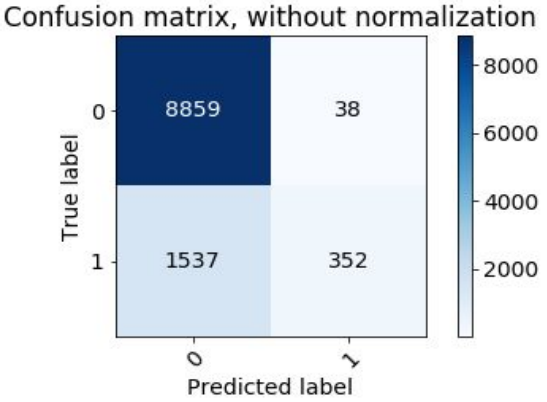
Churn has a high correlation with **TurnAutoRenewOff** and a moderate correlation with **number of previous churns**. It has a moderate negative correlation with **TenureMonth**, suggesting the longer a user has been a user, the less likely they are to churn.

EmailDomain, **Company**, and **country code** are string fields that have too many values to be useful as categorical variables and need to be grouped.

Build a model to predict churn and score subscribers that are coming up for renewal. Tell us how you expect your model to perform.

- I tested three models and discovered that the Random Forest model had the best predictive capability.
- The Random Forest model was able to correctly predict 19% of users who are about to churn while yielding almost no false positives.
- Consequently, I expect the model to identify every 1 in 5 users who are about to churn.





Describe how your model works, assuming that the audience is a non-technical marketing manager.

My model works by fitting a statistical model to the provided dataset which assigns a probability of churn to each customer.

The Random Forest model achieves this by identifying the input variables that best split the data into two homogenous groups of churners and non-churners.

Variables like AutoRenew and number of continuous months provide a good basis for predicting who is likely to churn.

What did you learn from your analysis and modeling about the characteristics of the subscribers most at risk of not renewing?

The subscribers most at-risk for not renewing are ones that have turned off their auto renew and/or don't have auto renew on. The number of continuous months the subscriber has been subscribed is also an important factor as well as whether they received a free trial.

The top 8 features of the model as identified by the Random Forest model are listed out on the right.

Feature	importance
TurnAutoRenewOff	0.299635
AutoRenewOn_no	0.181055
BeginningState_Renewal	0.103515
MaxContinuousActiveMonths	0.097124
TenureMonth	0.064875
StartTrial	0.063525
NActiveMonths	0.056777
NContinuousActiveMonths	0.046568

What recommendations would you make for the use of your model?

The model is usable as is because it does not produce many false-positives.

Based on the features identified it makes sense to tailor a retention strategy for customers who have adjusted their auto renew setting or have not set auto-renew on.

I would also recommend strategizing for new subscribers to encourage them to remain active for 4 months - a point at which the likelihood of churning is cut in half.