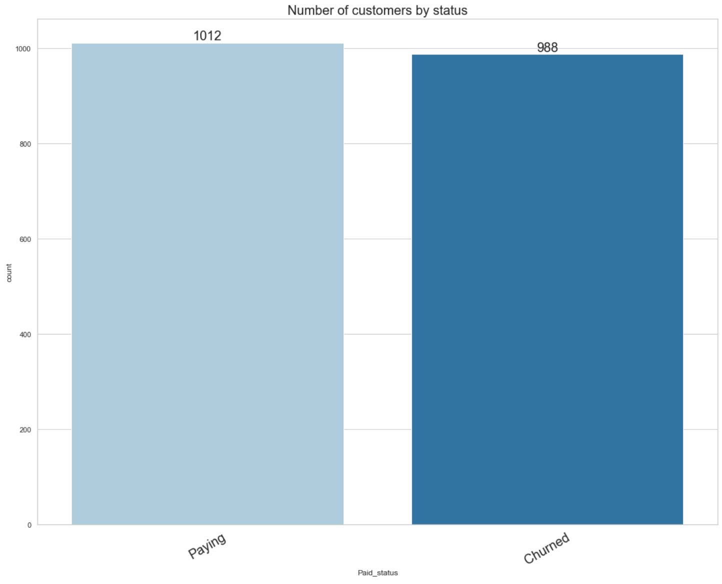
Customer Churn Prediction and Evaluation

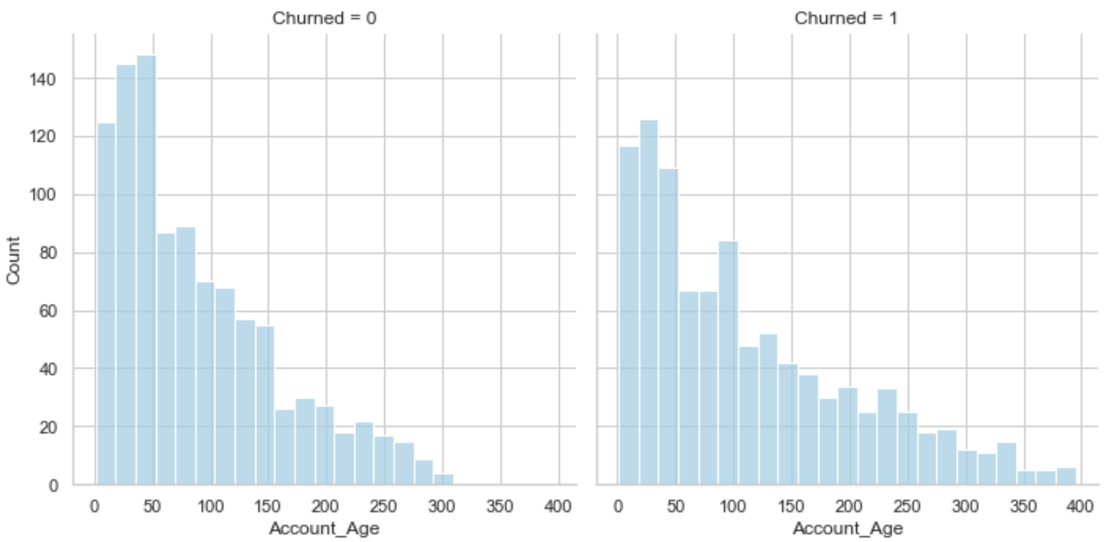
Business Audience

Agenda

- The state of churn
- High level analysis overview
- What causes churn?
- Recommendations and next steps

We observed a similar proportion of churned to paying customers with skewed account age distributions



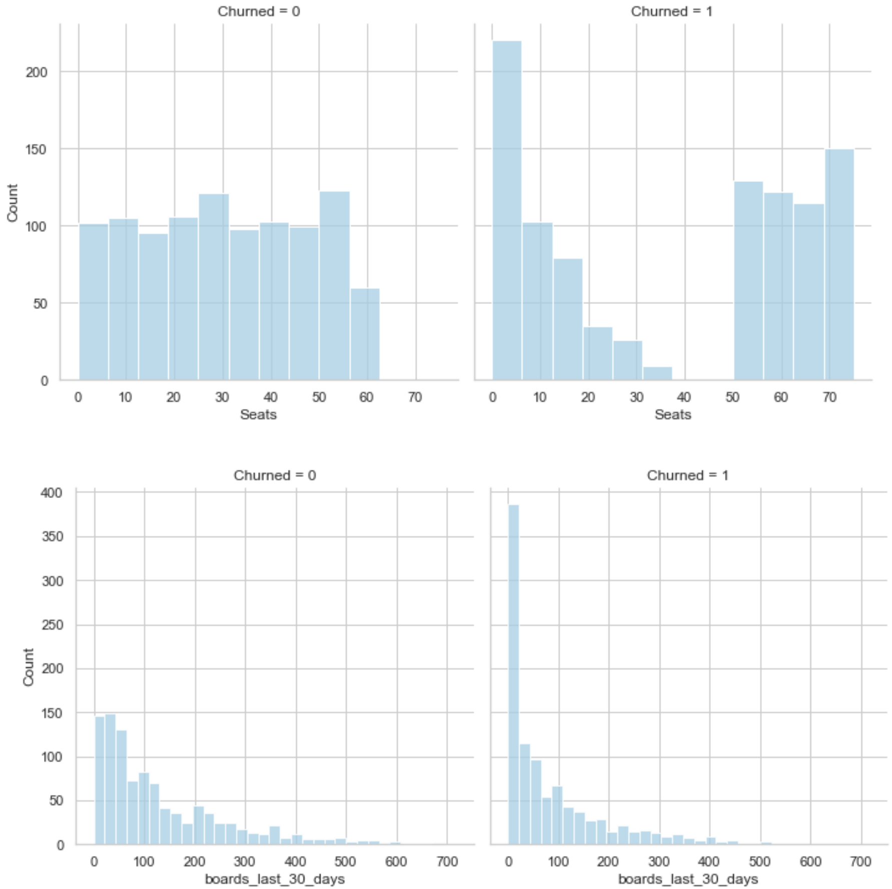


We analyzed 2000 randomly sampled accounts and various characteristics of each account (like usage, age, integrations, etc.)

Observations:

- The random sampling of accounts we analyzed was split 50.6% paying and 49.4% churned
- Churned accounts also have a higher likelihood of being older accounts by the time they churn

Account activity seems to be a good initial indicator of churn



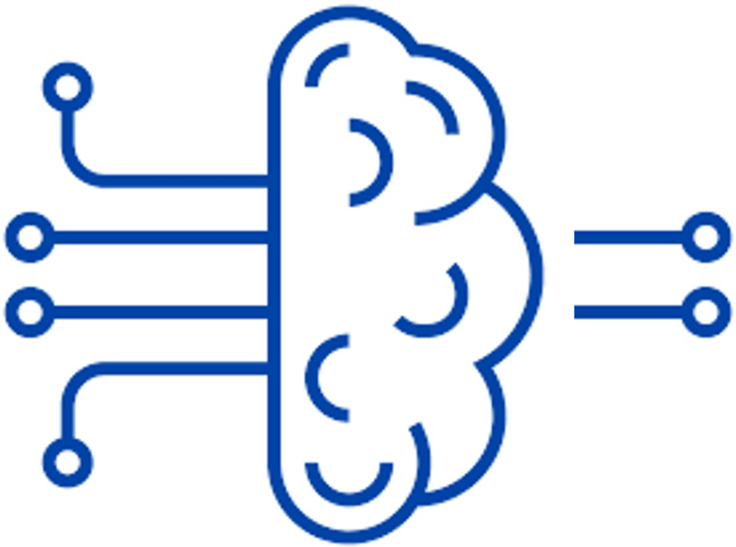
Observations:

- Churned accounts have a barbell distribution of accounts with a higher number of low seat numbers and high seat numbers
- Churned accounts also intuitively have fewer boards created in the last 30 days, while the opposite is true for Active accounts
- **CSMs should cultivate power users and power teams early in an account's history to help reduce churn.** As seen on the previous slide, older accounts were more likely to churn
    - Sales teams should be mindful of overselling into an account

We analyzed 2000 accounts to better understand what drives churn

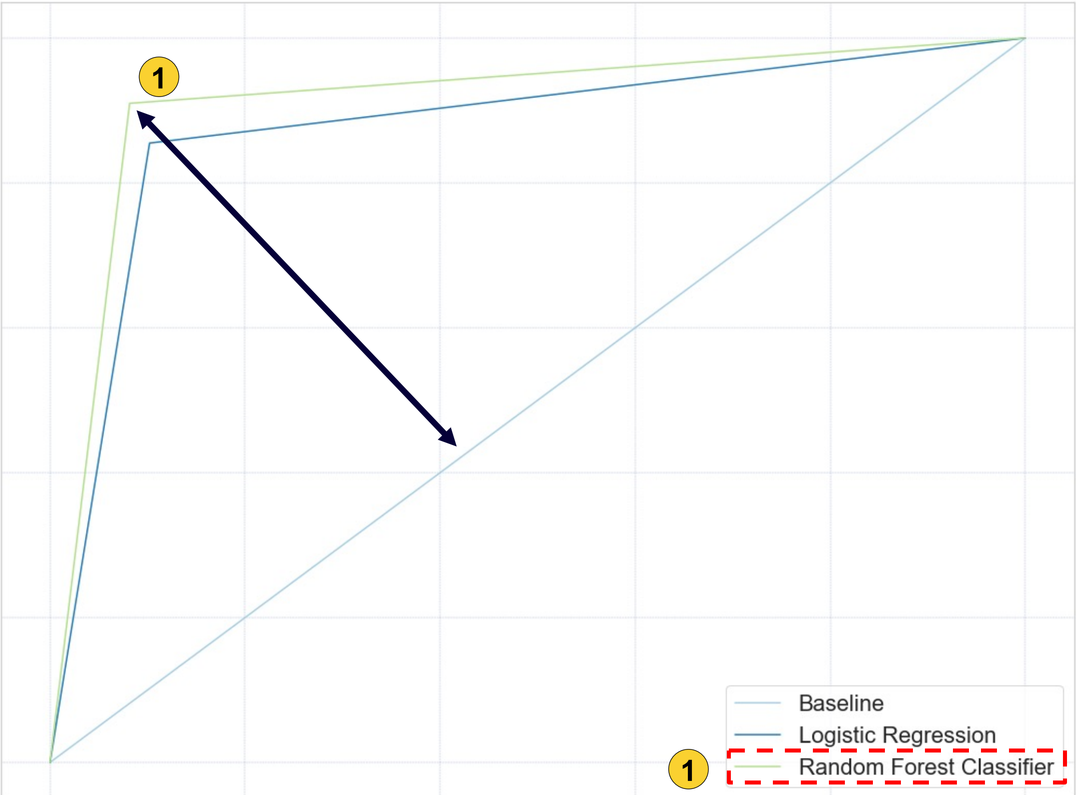
We took several data points about our customers and passed them through statistical models to produce a churn risk score for clients and to highlight the main drivers of account churn

- Account Age
- Templates_last_30_days
- Seat Licenses
- Boards Last 30 Days
- Projects Last 30 Days
- Avg. DAUs Last 30 Days
- Number of Integrations
- The Types of Integrations



Churn Risk Score
+
Major Churn Risk Factors

Using a statistical approach, our model greatly improves churn prediction



Observations:

- Predicting churn is a binary outcome with a near 50/50 probability when chosen at random (represented by the Baseline)

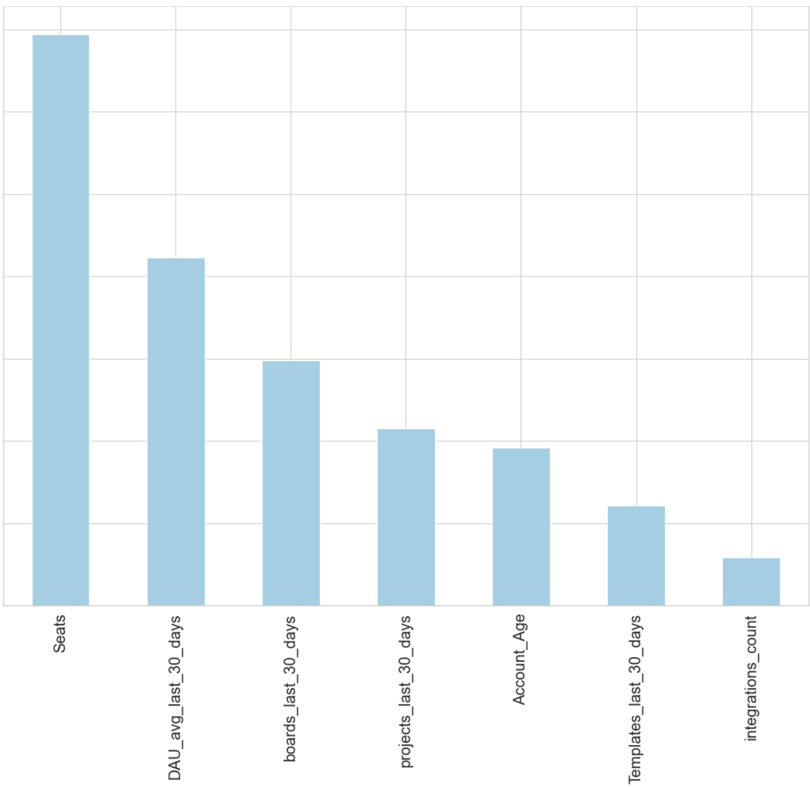
1 By utilizing statistical models, we can raise the accuracy of churn prediction from the Baseline to the Random Forest Classifier curve

  - All the area under the Random Forest Classifier curve is “alpha” our model can capture

- This model raises the accuracy of predicting churn from **50% (a random coin flip) to 91%**

What drives churn

Relative Importance for Churn Risk – Without Customer Integrations



Select Customer Attributes	Correlation with Churn
Account_Age	14%
Seats	15%
boards_last_30_days	-21%
projects_last_30_days	-17%
DAU_avg_last_30_days	-18%

Observations: The most important factors for churn (relatively, not directionally) can be seen in the graph to the left

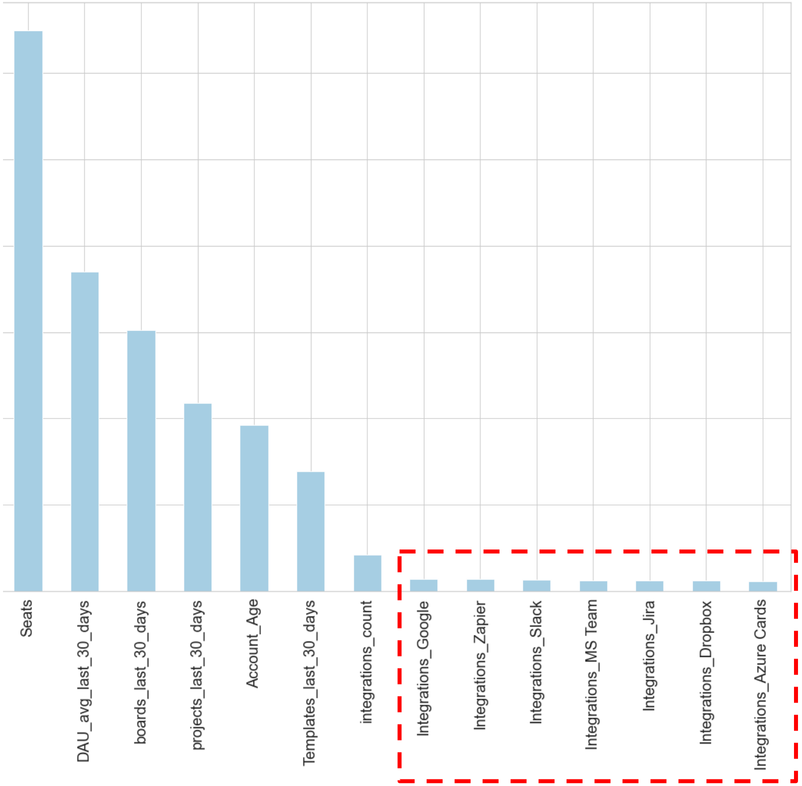
- Seats: Positive correlation between churn and seats. Customers may have been oversold seat numbers and will churn if they don't see positive value for this large line item

- Usage (projects, boards, DAUs): Negative correlation between usage and churn. This is intuitive and should reinforce the importance of CSMs fostering healthy product usage

- Account Age: A higher number of accounts churned who had longer tenures. CSMs and Product Marketing should potentially be more aggressive in letting older accounts know how the product has improved

Surprisingly, the types of integrations didn't matter much in predicting churn

Relative Importance for Churn Risk – With Customer Integrations



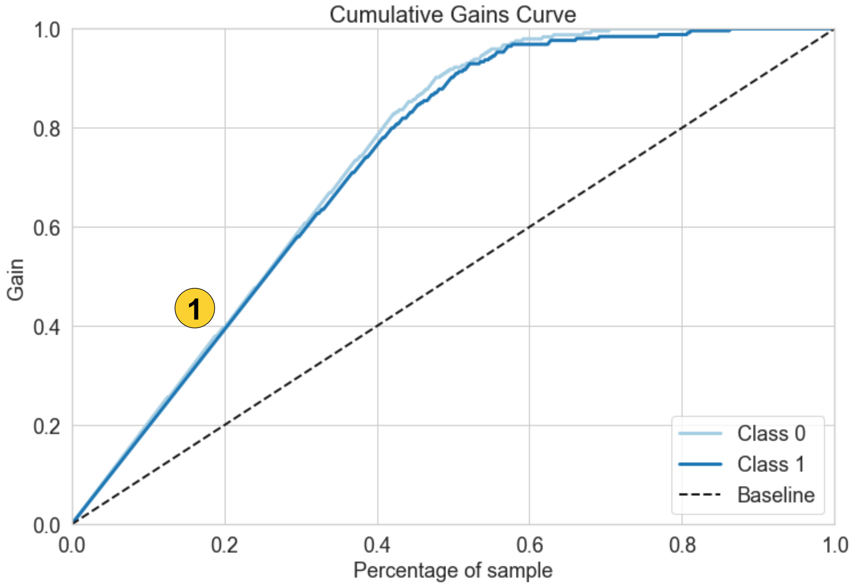
Observations:

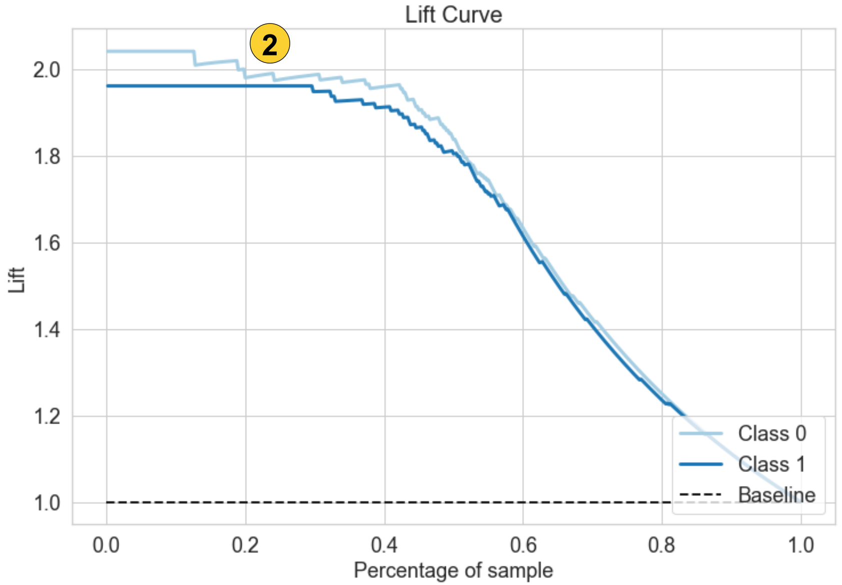
- Highlighting the types of integrations the Company offers may be helpful pre-sale, but post-sale it seems to have little impact on how sticky the product is

Recommendations:

- CSMs post-sale should consider not focusing on any specific integration unless the customer specifically inquires

How many customers should we reach out to about their churn risk?





1 By selecting 20% of top model ranked customers we would expect to see 40% of churning customers

- Likewise, by selecting 60% of top model ranked customers we would see ~100% of churning customers

2 By picking 20% of customer, this model is ~2x better at finding churn risk customers than picking at random

- The Gains Chart can help determine how many customers to sample and target for churn prediction based on cost and expected revenue, i.e., **how can we prevent as much churn as possible by reaching out to the fewest number of customers**

Recommendations to lower churn

Recommendations and Next Steps:

1. Work with CSMs to check in with accounts at greatest risk of churn. Additionally, forecast how much time should be dedicated to this task for optimal churn prevention

2. Work with product and engineering to see if we can get Heap (or any engagement) data and see what hurdles exist for users creating boards and using the product. Where is product drop-off occurring?

3. Post-sale CSMs should focus on creating as many active users as possible rather than technical integrations

4. Number of DAUs seems much more important than the number of seats. Consider creating financial incentives for CSMs for hitting client DAU goals and maybe deemphasize Seat goals

5. Given older accounts propensity to churn, as accounts age make sure to follow up more frequently with product updates and improvements

6. Work to deploy this model and integrate churn risk scores into Salesforce

    1. Weight MRR by churn risk to flag high risk dollar weighted accounts

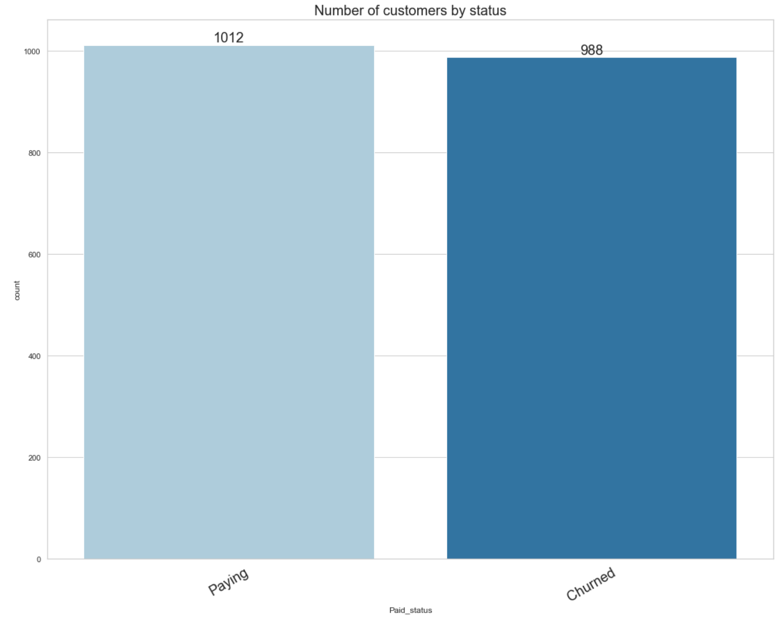
Analytics Audience

Agenda

- Visual data inspection and cleaning
- Exploratory data analysis
- Model selection
- Model output
- Recommendations and next steps

Data inspection of 2000 randomly selected accounts

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 10 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   AccountID               2000 non-null   int64
 1   Account_Age             2000 non-null   int64
 2   Seats                   2000 non-null   int64
 3   Paid_status             2000 non-null   object
 4   boards_last_30_days     2000 non-null   int64
 5   projects_last_30_days   2000 non-null   int64
 6   DAU_avg_last_30_days    2000 non-null   int64
 7   integrations_count      2000 non-null   int64
 8   Integrations_names      1621 non-null   object
 9   Templates_last_30_days  2000 non-null   int64
dtypes: int64(8), object(2)
memory usage: 156.4+ KB
```

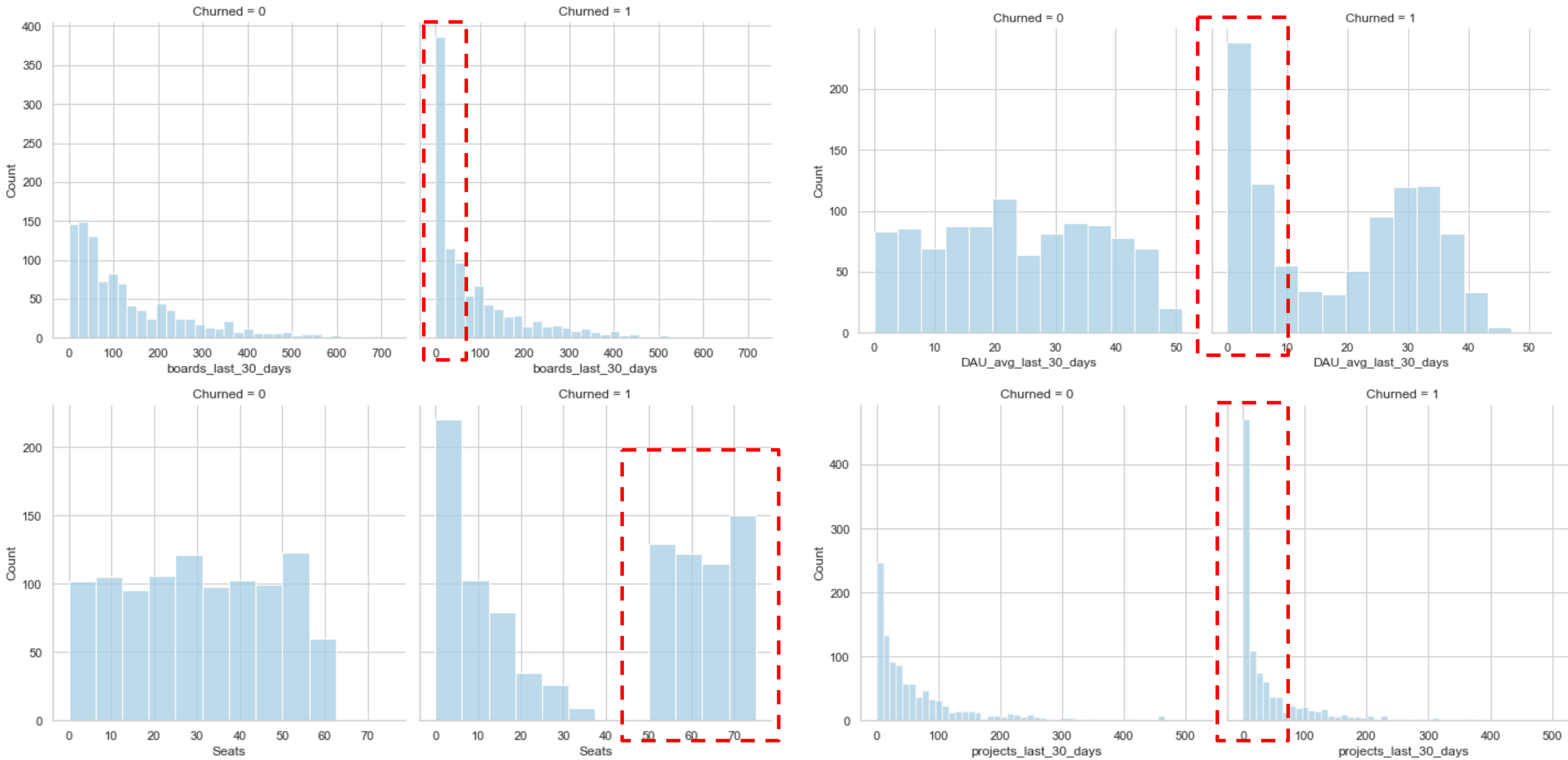


Observations:

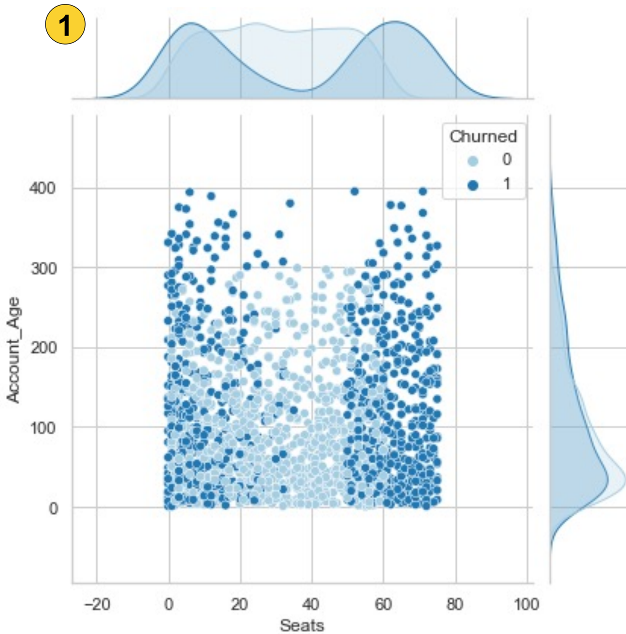
- The random sampling of accounts we analyzed was split 50.6% paying and 49.4% churned
- For modeling purposes dummy variables were created for integration names (or types) and AccountID was dropped
- Churned customers are represented by 1 in the model while active customers are 0

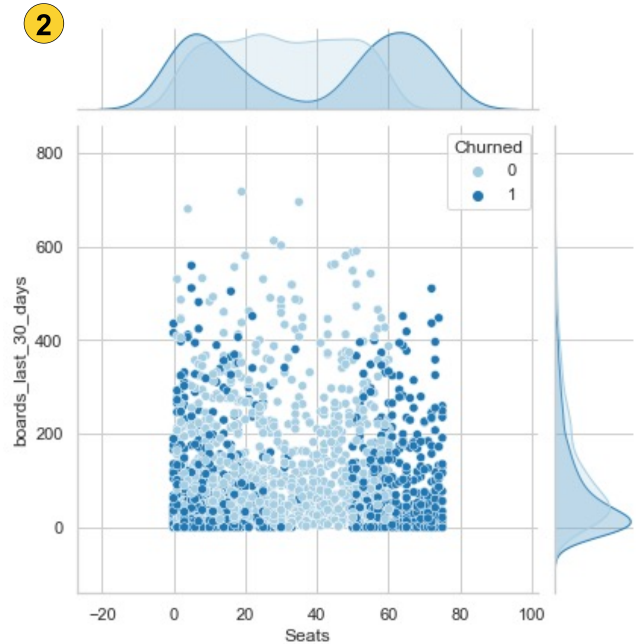
Exploratory data analysis

Lower product usage customers have high propensity to churn, but surprisingly high seat customers do as well



Exploratory data analysis (cont.)





Observations:

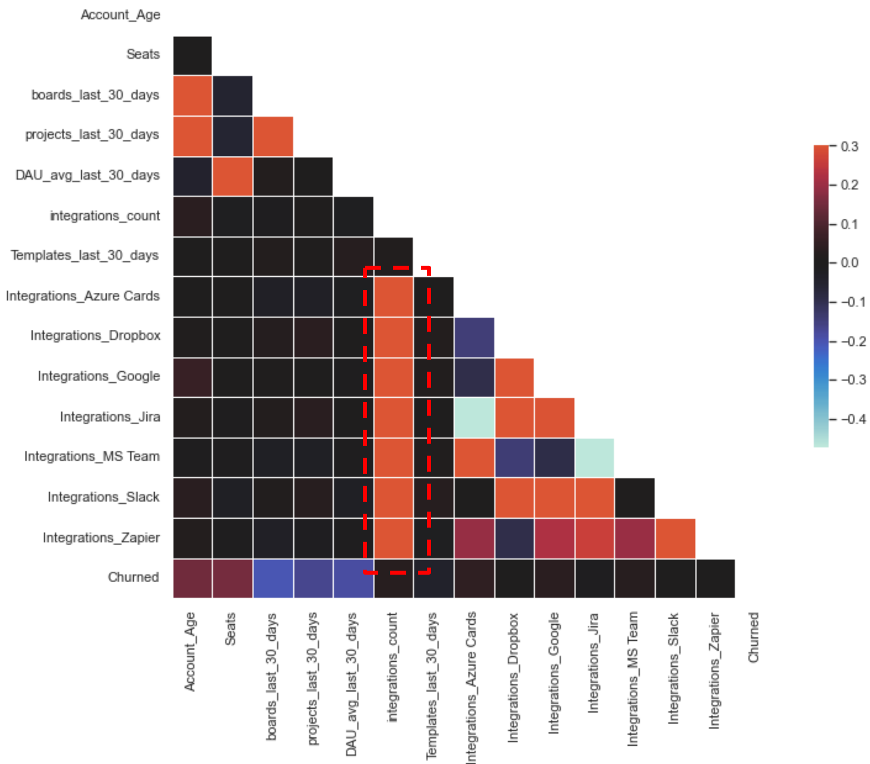
1. Churned accounts have a high distribution of low seat and high seat numbers, fatter tail distribution on account age with a higher concentration of older + high seat accounts

2. Active accounts on the other hand have a fatter tail distribution on boards created over the last 30 days

- Cultivating power users and power teams early in an account's history is likely important for retention

- Sales should be mindful of overselling into an account

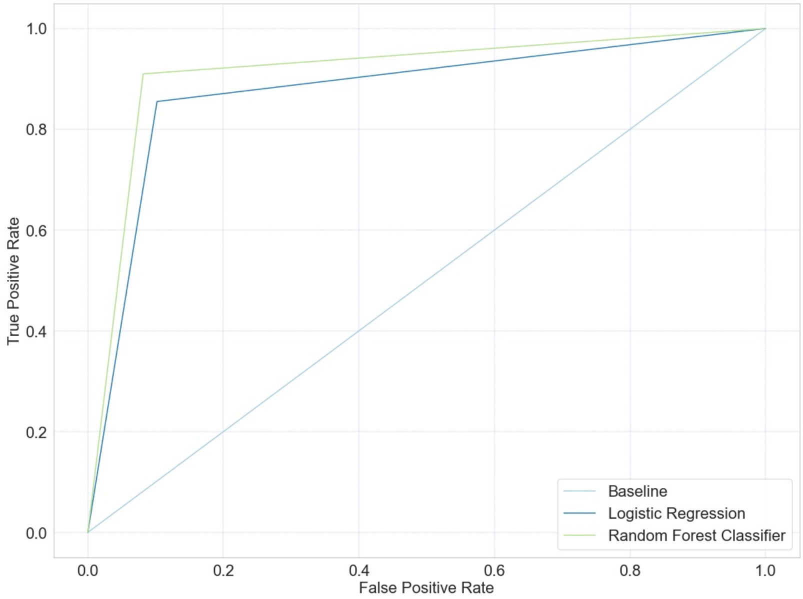
Exploratory data analysis – multicollinearity



	variables	VIF
0	Account_Age	1.910538
1	Seats	5.328353
2	boards_last_30_days	3.264756
3	projects_last_30_days	2.414500
4	DAU_avg_last_30_days	5.331882
5	integrations_count	inf
6	Templates_last_30_days	1.004259
7	Integrations_Azure Cards	inf
8	Integrations_Dropbox	inf
9	Integrations_Google	inf
10	Integrations_Jira	inf
11	Integrations_MS Team	inf
12	Integrations_Slack	inf
13	Integrations_Zapier	inf

- Several independent variables are correlated, and some have infinite VIF (variance inflation factor) scores
- The integration types are correlated with integration count which intuitively makes sense and will likely impact the model
- We will try modeling churn prediction first using these correlated variables, but later removing them to see if model accuracy improves
- Seats and DAU VIFs are barely above 5 and below 10 so those independent variables will stay in the model for now (despite their high correlation)

Using a Random forest model provides higher accuracy than a logistic regression and baseline



Random Forest Results

	precision	recall	f1-score	support
0	0.89	0.91	0.90	245
1	0.91	0.89	0.90	255
accuracy			0.90	500
macro avg	0.90	0.90	0.90	500
weighted avg	0.90	0.90	0.90	500

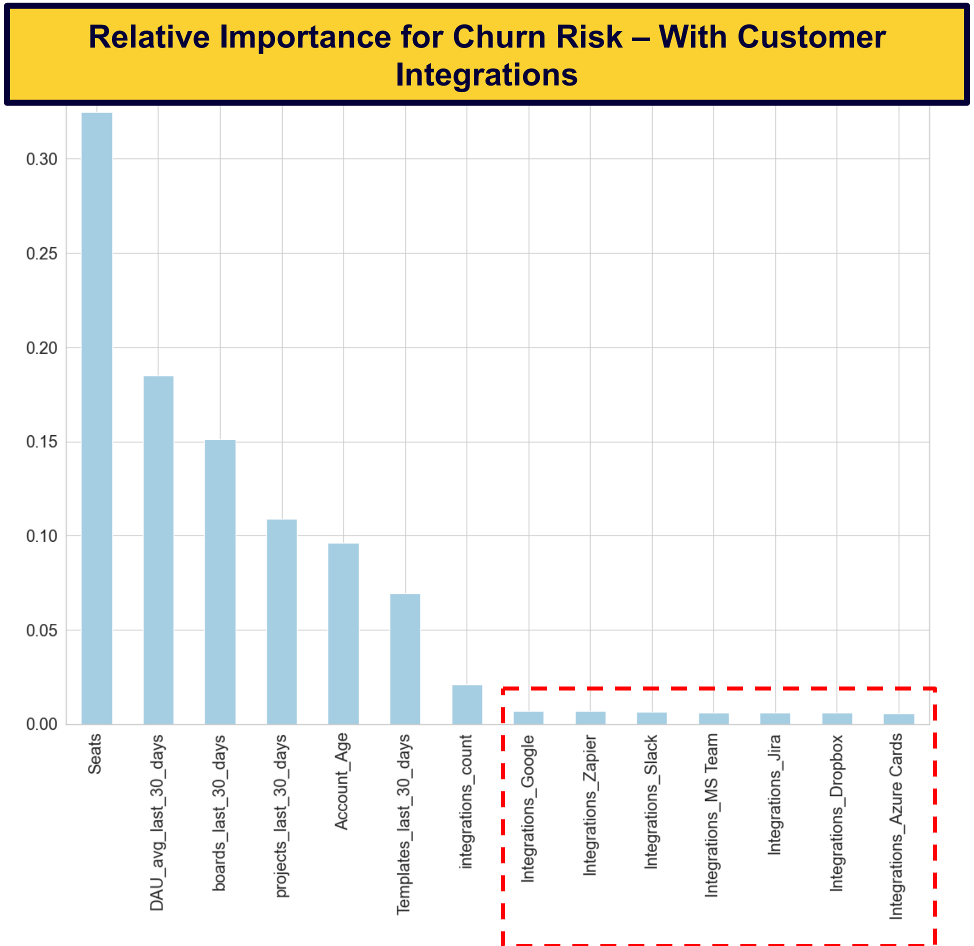
Logistic Regression Results

	precision	recall	f1-score	support
0	0.86	0.90	0.88	245
1	0.90	0.85	0.88	255
accuracy			0.88	500
macro avg	0.88	0.88	0.88	500
weighted avg	0.88	0.88	0.88	500

Why a Random Forest Model?:

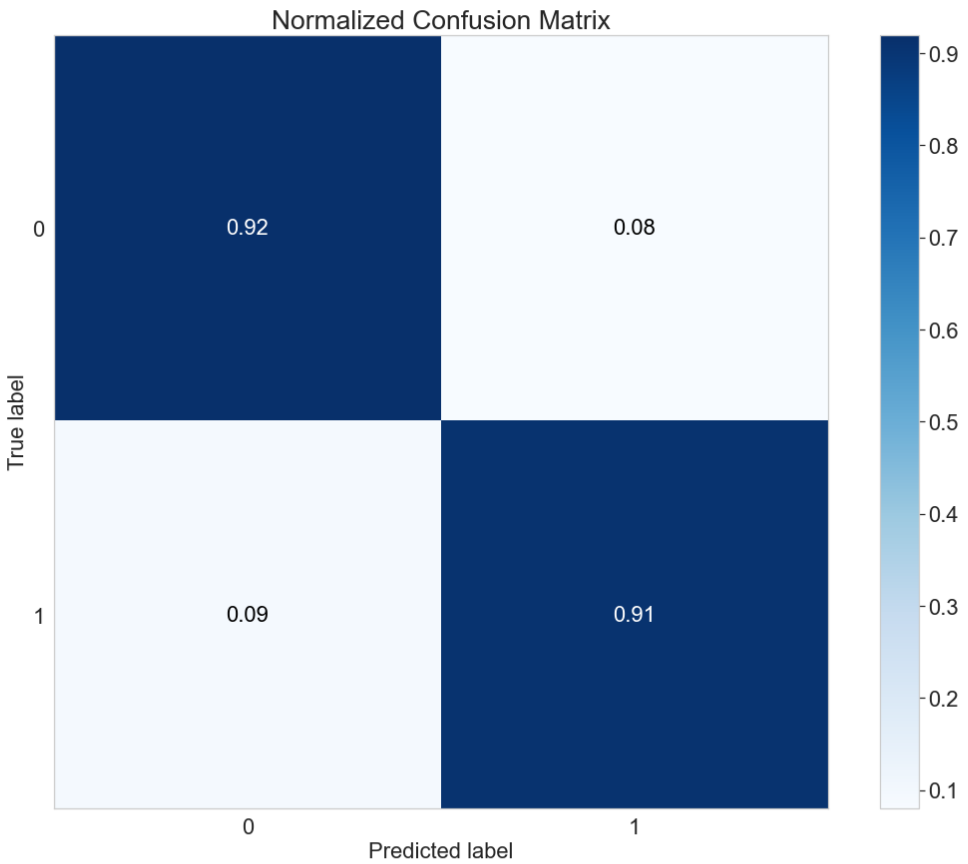
- Random forest models help reduce overfitting by creating random subsets of the features and building smaller trees using those subsets
- While random forest models can be slow, prediction speed isn't as important for this analysis
- Some of the features seem to be non-linear
- The model gives a good indicator of the importance assigned to features
    - Although the model lacks in some of the explanatory power and directions of these features. Looking at correlations and descriptive statistics can help lessen this downside

Integration types have high VIF scores and low feature importance



- As seen on previous slides, the types of customer integrations may be a drag on the model
    - These independent variables have infinite VIF scores and are very correlated with integration count
- Running a random forest model also places very low importance on integration types as independent variables
- Removing them should improve model accuracy and the integration count variable should capture enough information about integrations

Random forest performance without integration types



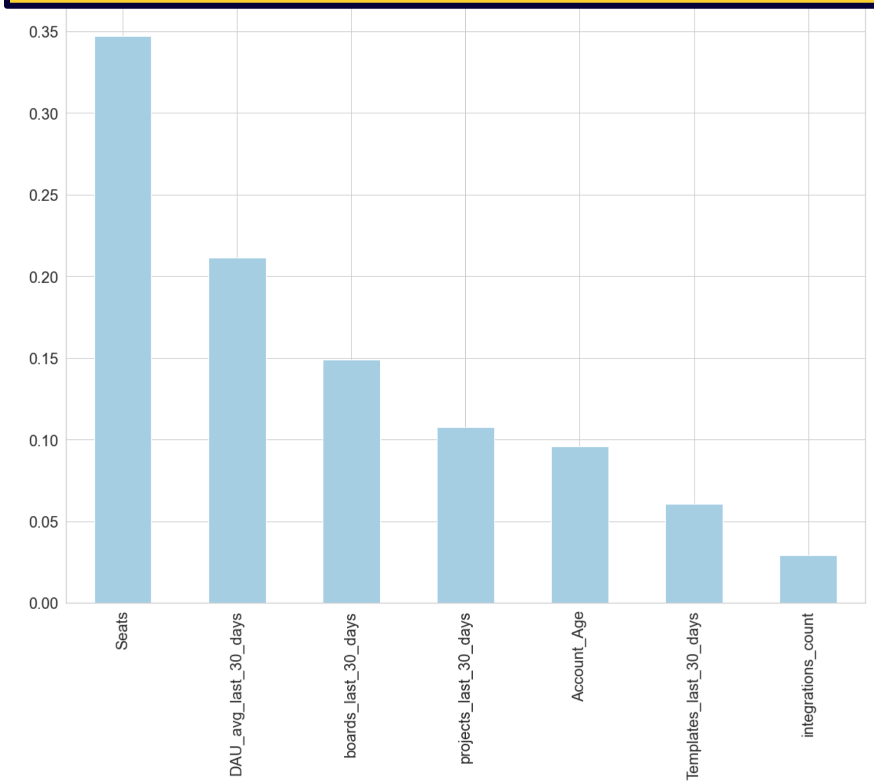
Random Forest Results

	precision	recall	f1-score	support
0	0.91	0.92	0.91	245
1	0.92	0.91	0.92	255
accuracy			0.91	500
macro avg	0.91	0.91	0.91	500
weighted avg	0.91	0.91	0.91	500

- The model’s accuracy and confusion matrix improves when eliminating integration types
  - Accuracy improved +1% and F1-score +2%
- Going forward, the preferred model for predicting churn will be a **random forest classifier that does not include the integration types as independent variables**

Model implications on what is important for predicting churn

Relative Importance for Churn Risk – Without Customer Integrations



The values on the left were computed as the mean and standard deviation of accumulation of the impurity decrease within each tree. This method was primarily deployed for simplicity and speed

- Seats: Positive correlation between churn and seats. Customers may have been oversold seat numbers and will churn if they don't see positive value

- Usage (projects, boards, DAUs): Negative correlation between usage and churn. This is intuitive and should reinforce the importance of CSMs fostering healthy product usage

- Account Age: A higher number of accounts churned who had longer tenures. CSMs and Product Marketing should potentially be more aggressive in letting older accounts know how the product has improved

Limitations: feature importance is biased towards features with high cardinality and correlated features with be given similar importance

Recommendations to lower churn

Recommendations and Next Steps:

1. Work with CSMs to check in with accounts at greatest risk of churn
2. Work with product and engineering to see if we can get Heap (or any engagement) data and see what hurdles exist for users creating boards and using the product. Where is product drop-off occurring?
3. Post-sale CSMs should focus on creating as many active users as possible rather than technical integrations
4. Number of DAUs seems much more important than the number of seats. Consider creating financial incentives for CSMs for hitting client DAU goals and maybe deemphasize Seat goals
5. Given older accounts propensity to churn, as accounts age make sure to follow up more frequently with product updates and improvements
6. Work to deploy this model and integrate churn risk scores into Salesforce
    1. Weight MRR by churn risk to flag high risk dollar weighted accounts
    2. Before deploying, continue tuning the model, testing new independent variables to improve model performance, and testing against new incoming data