Predict In-App Purchases

Evan Calkins
Ran Huang
Alan Perry

Agenda

- Dataset & Research Question
- EDA
- Data Prep & Features
- Modeling & Results
- Business Insights
- Things Learned

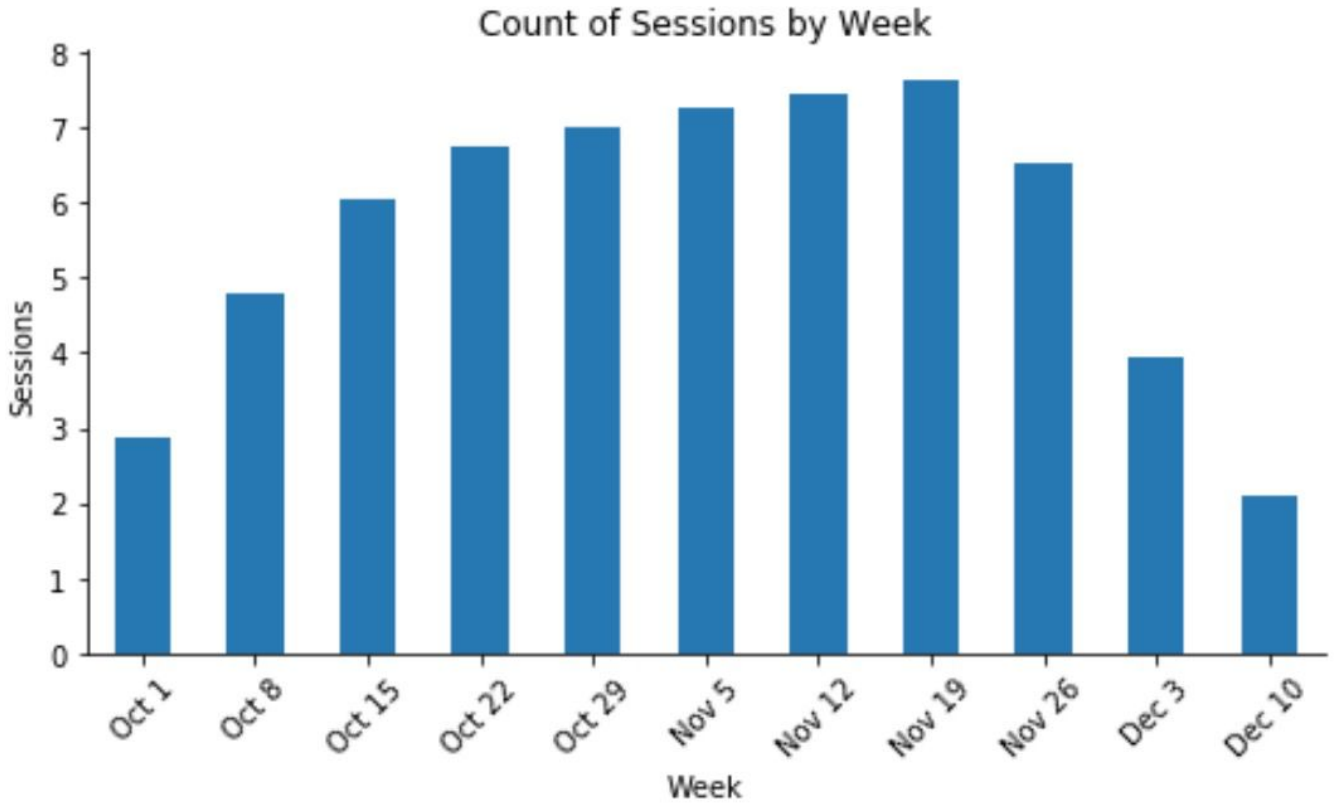
Dataset & Research Question

- Dataset provided by Leanplum
- App data on users created between 10/1/2018 and 11/30/2018
  - Sessions
  - Events
  - Attributes
- Goal is to predict whether or not a user will purchase within the next 7 and 14 days

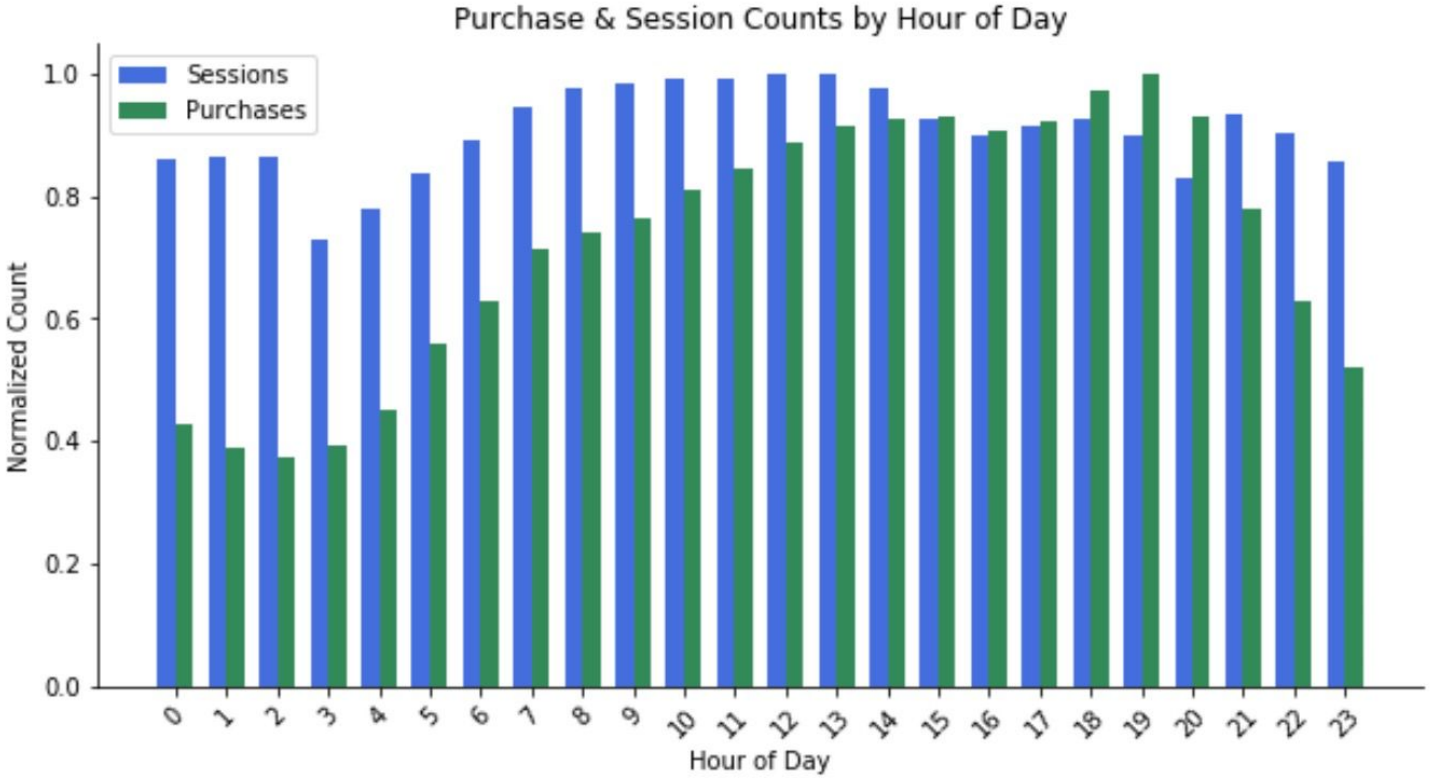
Dataset & Research Question

- Sessions
  - Time series dataset
  - Tracks when a user was on the app
- Events
  - What the user did on the app (e.g. purchase, swipe)
  - Limited understanding due to confidentiality
- Attributes
  - State of the user at the time of the session (e.g. churn score)
  - Mostly numerical
  - Limited understanding due to confidentiality

EDA

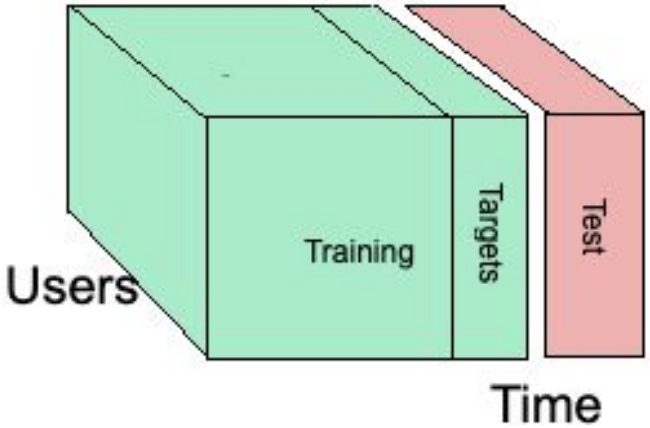

Count of Sessions by Week
Sessions
0
1
2
3
4
5
6
7
8
Oct 1
Oct 8
Oct 15
Oct 22
Oct 29
Nov 5
Nov 12
Nov 19
Nov 26
Dec 3
Dec 10
Week

EDA


Purchase & Session Counts by Hour of Day
Sessions
Purchases
Normalized Count
0.0
0.2
0.4
0.6
0.8
1.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Day

Data Prep & Features

- Create target labels using most recent two weeks of data
  - User purchased
  - User did not purchase
- Compute train features using prior two months of data
- Compute test features using full dataset



Data Prep & Features

- Initial features computed using data that were well understood
  - Past purchases
  - Past sessions
- Created features based on frequency and recency
  - Count features
  - Sum features
  - Last week, last 2 weeks, and total

Data Prep & Features

- Additional features created using lesser understood data
  - Other events (non-purchases)
  - Attributes
- Computed correlation of features with target variable to determine which to send to our model

	label	e1_count	e5_count
label	1	0.0799583	0.178831
e1_count	0.0799583	1	0.822344
e5_count	0.178831	0.822344	1
e6_count	0.166512	0.808274	0.995402
e14_count	0.100155	0.736403	0.76054
e4_count	0.165837	0.766358	0.984073
e40_count	0.156778	0.730259	0.948474
e7_count	0.266991	0.248757	0.523259
e41_count	0.128703	0.696752	0.862381
e3_count	0.127246	0.693115	0.860149
e42_count	0.123703	0.68794	0.857907

Data Prep & Features

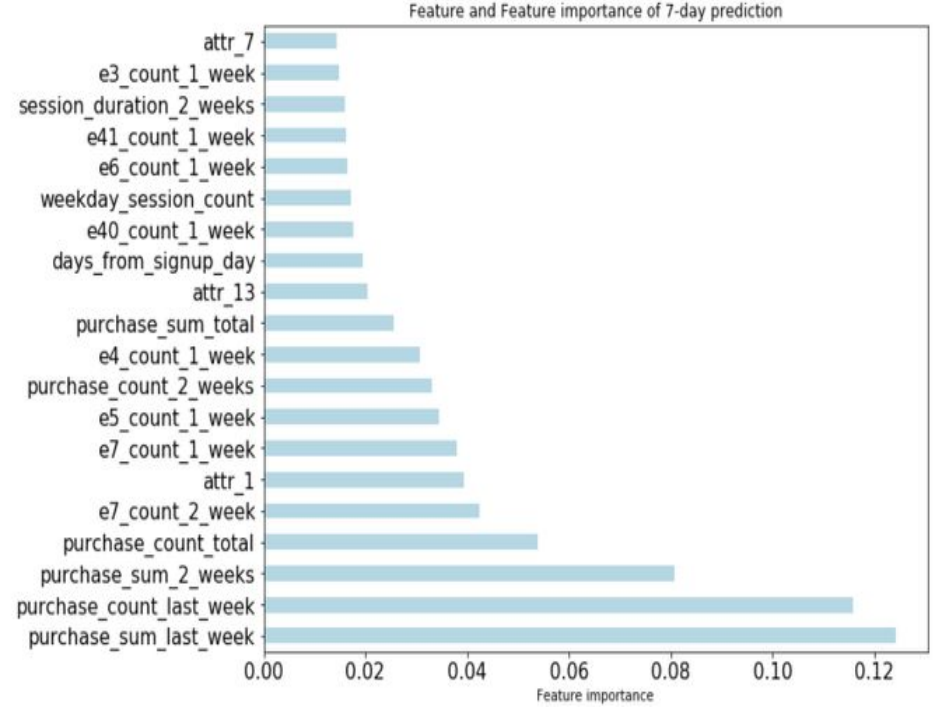
- Attributes 4 and 9 contained text data specific to the app
- Generated additional features by extracting text data
  - Vader sentiment score
  - Average word length

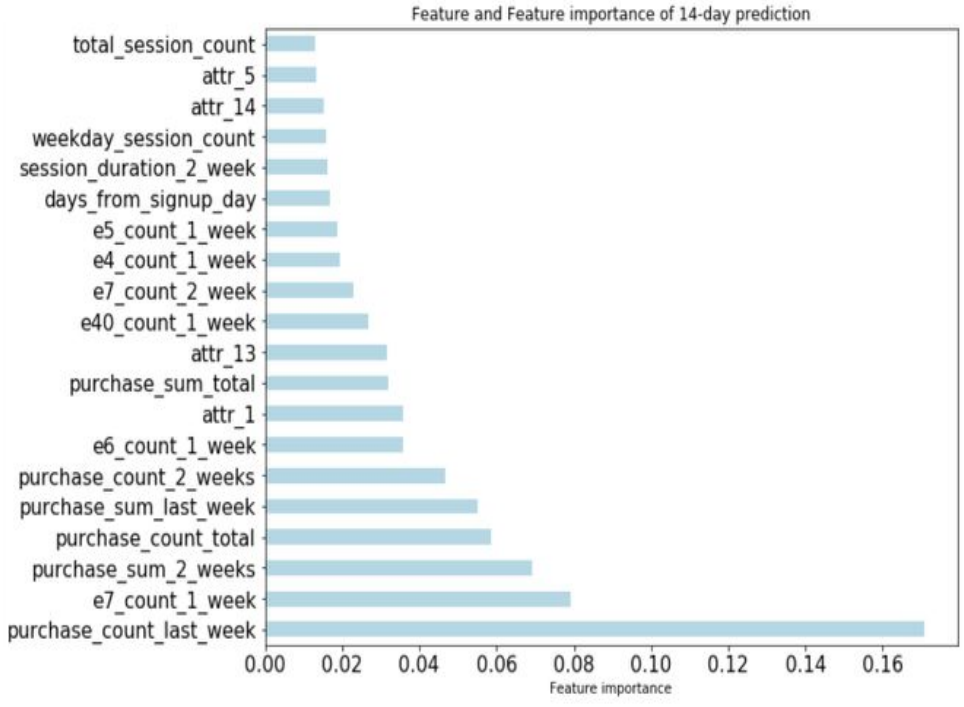
Modeling & Results

AUC score of validation set and test set on Kaggle for each model

Algorithm(s)	Datasets	Val AUC (7 day)	Val AUC (14 day)	Kaggle Score
XGBoost	Event, Session	0.96695	0.95666	0.9443
RF	Event, Session	0.973	0.964	0.96159
XGBoost	Event, Session, Attribute	0.975857	0.968354	0.96936
RF	Event, Session, Attribute	0.975009	0.967081	0.96804
Stacking	Event, Session, Attribute	0.976246	0.968574	0.96894

Modeling & Results





Business Insights

- Most helpful features were related to past purchases
  - Keep your paying customers happy
- Event 7 is a precursor to user purchasing
  - Increase Event 7 ➔ Increase Purchases
- Soft predictions can be used to rank users by likelihood of purchasing
  - Segment users and focus marketing efforts on middle ~1/3

Things Learned

- Feature engineering can have much greater impact than model selection / hyperparameter tuning
- The most important features are often the most logical
- If your data is too big to work with in Pandas, try reducing the size first before resorting to big data technologies

Thank you!