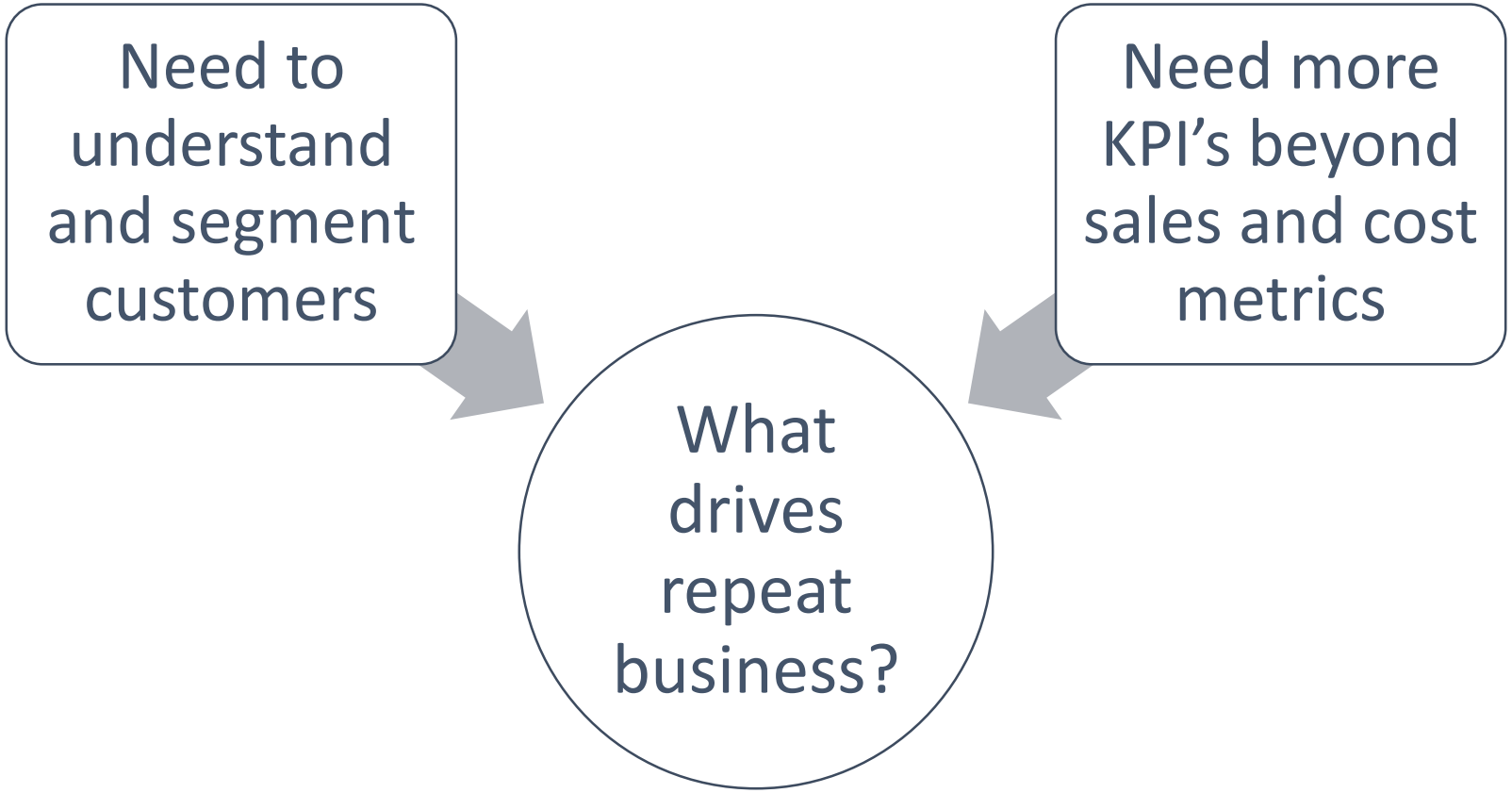
repeataly

predicting repeat customers with sales data

Zen Yui | 2016-03-07

Eataly needs to better analyze its customers



problem | EDA | models

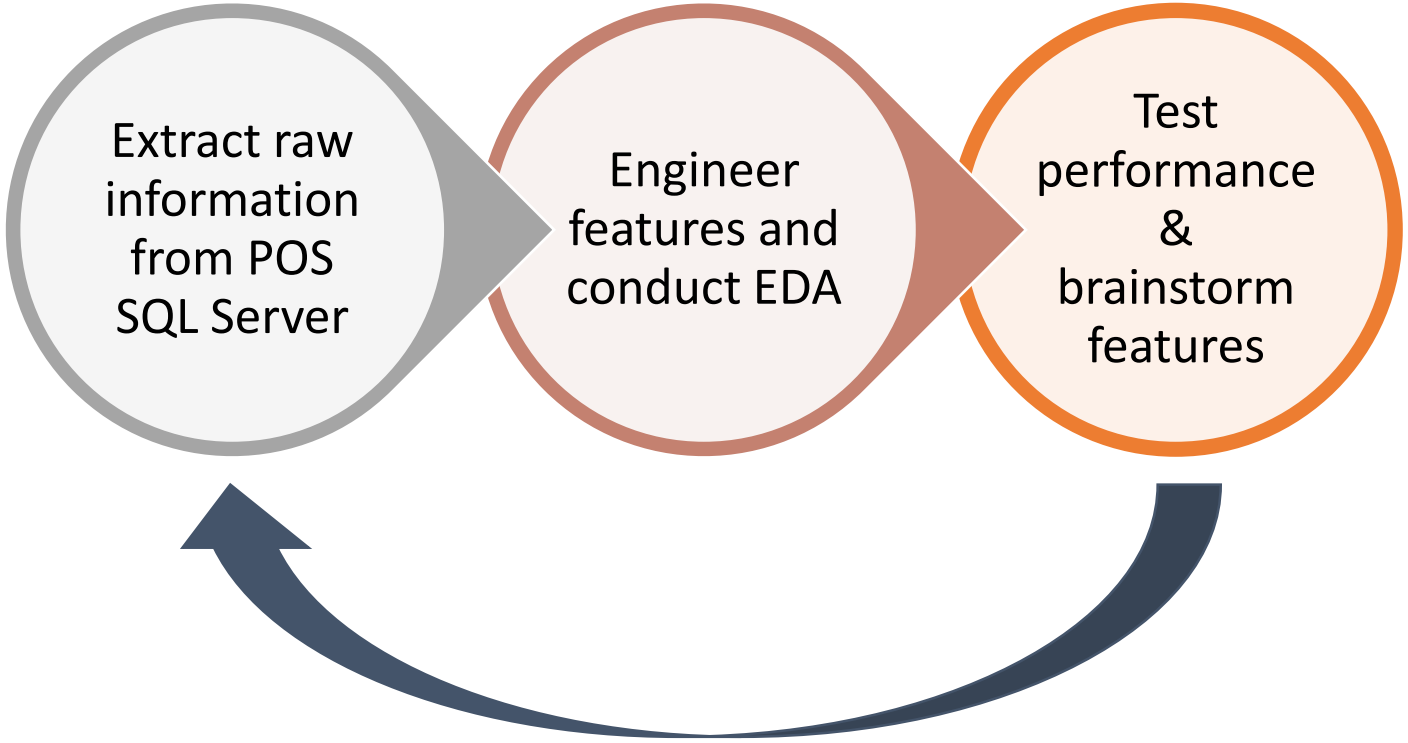
repeataly

The POS should contain enough data to make predictions

Question: *What features and values from Eataly's POS transaction data will best predict the likelihood of a repeat visit within 90 days?*

Hypothesis: *Prior visit count will likely prove to be the most important feature, but other features from "domain knowledge" should give an interesting perspective on customer behavior.*

The project followed a 3-step process



problem → EDA → models

Raw data came from two queries of different grains

Query 1:

Ticket-specific features that, in 1 row, summarize the retail as many customer behavior concepts as possible. This data is meant for clustering.

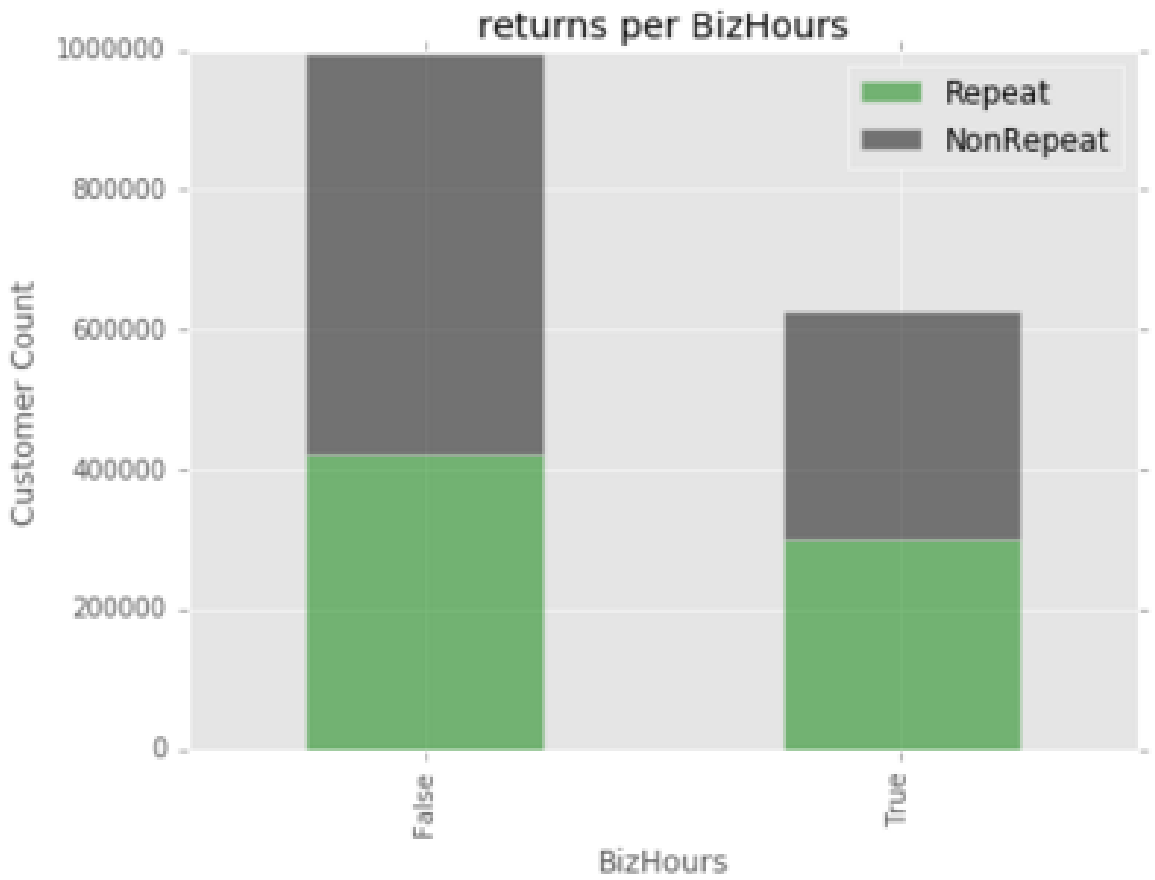
Query 2:

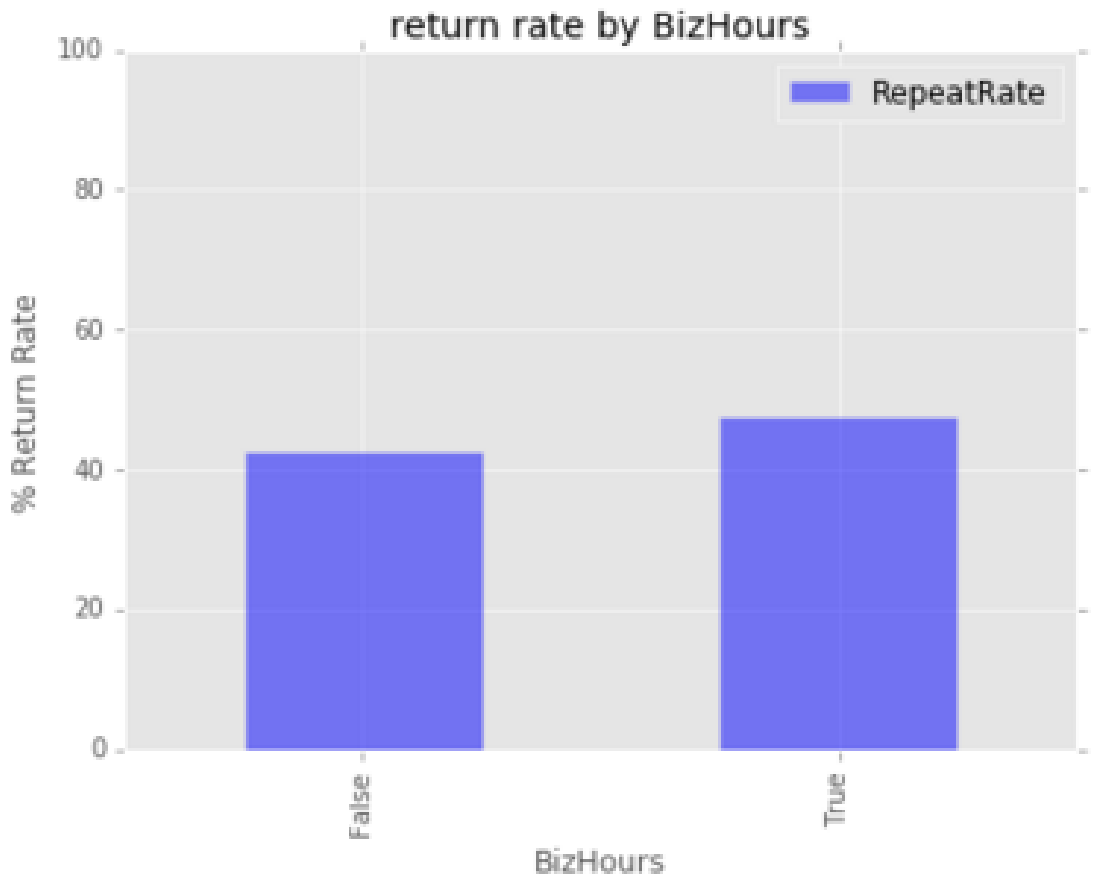
Hourly Ticket count per day to summarize how busy the store is, as Eataly has a negative reputation for traffic in peak hours

EDA was difficult given feature correlation

data dictionary		
Feature	**Type**	**Description**
TicketDate	datetime	Date of purchase
TicketTime	float	Decimal representation of hour of checkout
PriorVisits	int	Number of visits prior to current ticket
GiftCardLines	int	Lines of gift cards in basket
OilLines	int	Lines of oil products in basket
RotisserieLines	int	Lines of rotisserie products in basket
NetAmount	float	Net spend in dollars
DiscountAmount	float	Net discount in dollars
UniqueItems	int	Count of unique products in basket
UniqueCategories	int	Count of unique categories in basket
ReturnedBags	boolean	Boolean indicator of bag return for refund
TopItemLines	int	Count of top 100 repeat-customer products
RepeatProducts	int	Count of items purchased both last and current visit
TicketMonthDecimal	float	Decimal representation of month and day
WeekdayNumber	int	Integer weekday (1=Monday…7=Sunday)
IsFrontEnd	boolean	True/False indicator that customer used main exit
FreshLines	int	Count of fresh prodcut lines in basket
RetailSpendCnt	float	Locally centered spend on retail products
QsrSpendCnt	float	Locally centered spend on QSR products
Ticketcount	int	Count of tickets happing in that hour
TotalLines	int	Total sale lines and return lines in ticket
ProductVariety	float	Unique products divdied by whole basket
BizHours	boolean	True/False indicator of shopping weekdays before 5pm
WillReturn	boolean	True/False prediction class for repeat visit

BizHours was not a clear predictor!



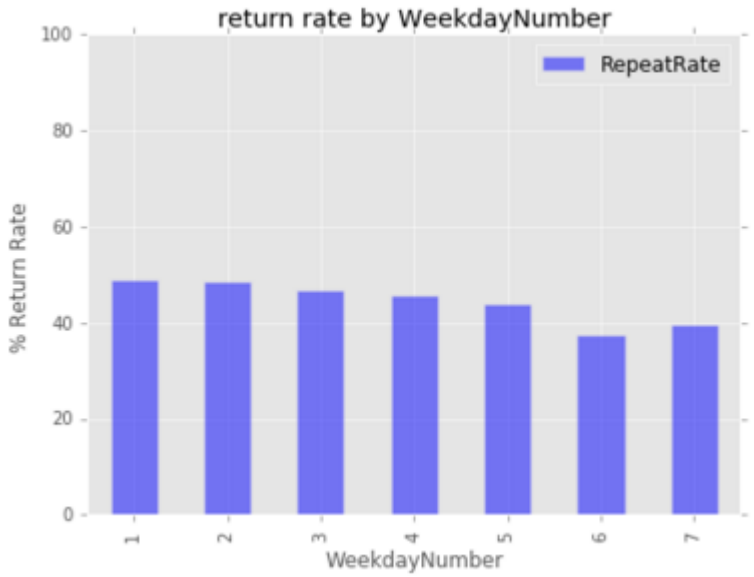


problem → EDA → models

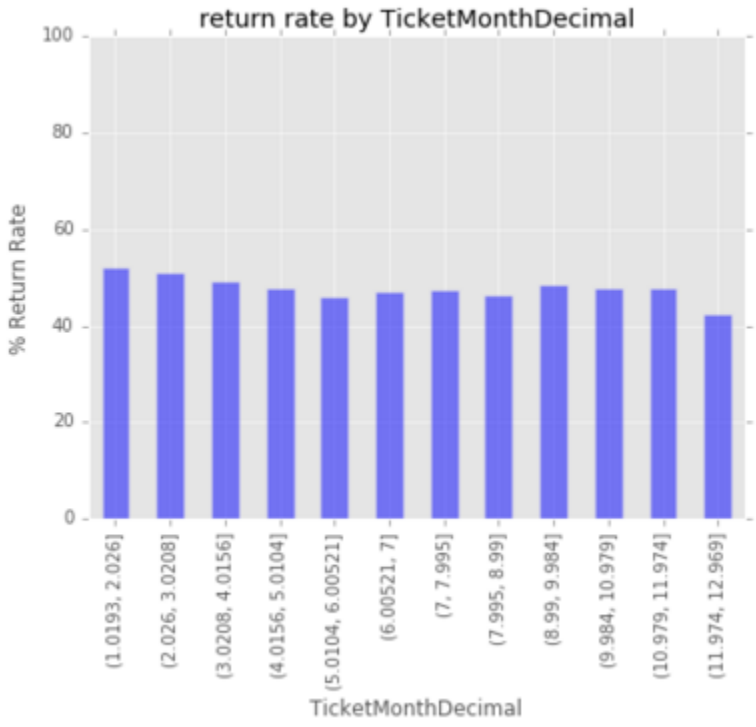
repeataly

Time series features were not as predictive as hoped

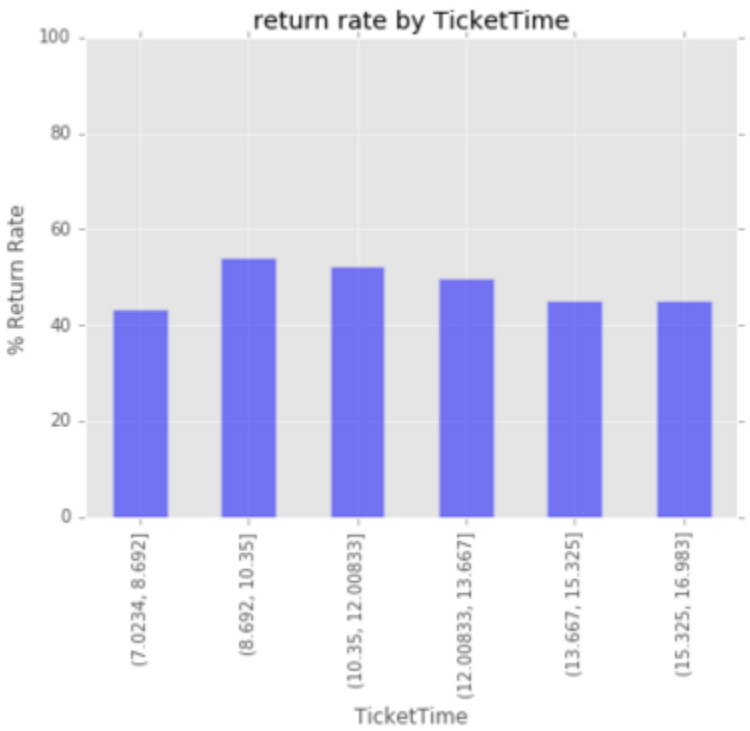
WEEKDAY



MONTH



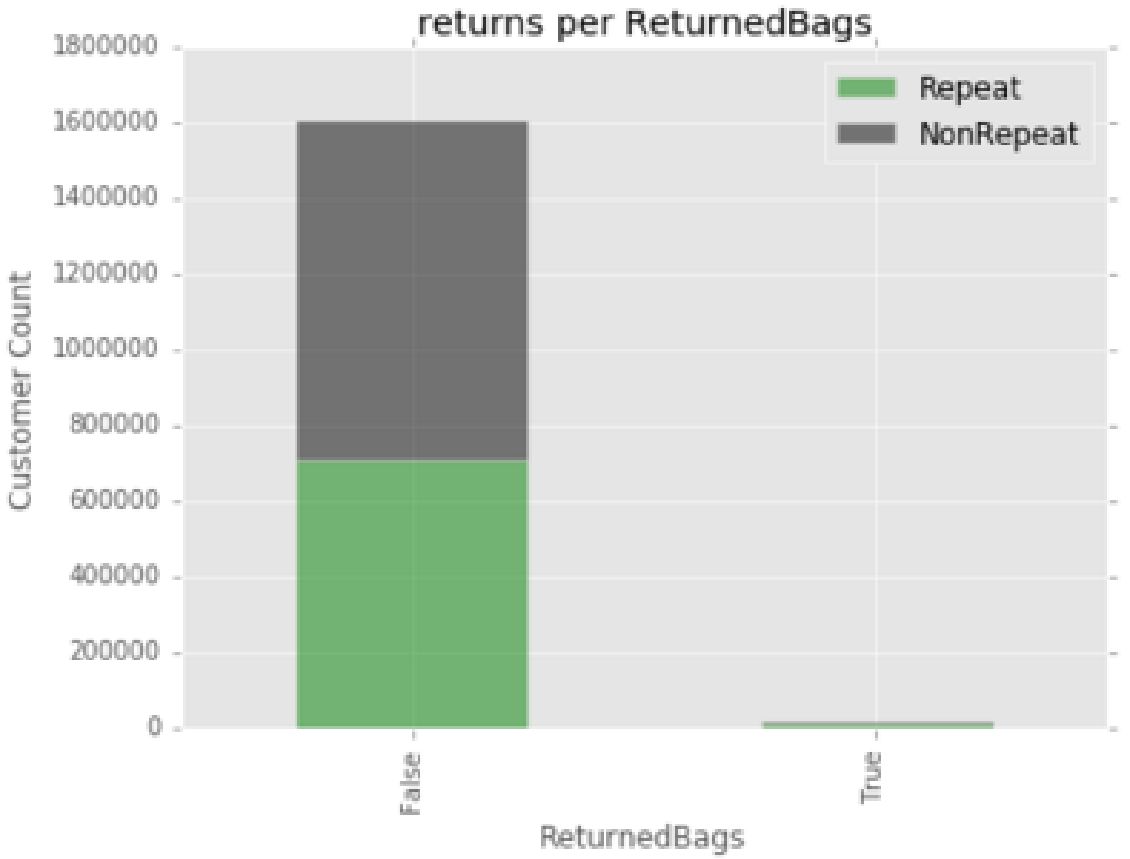
TIME

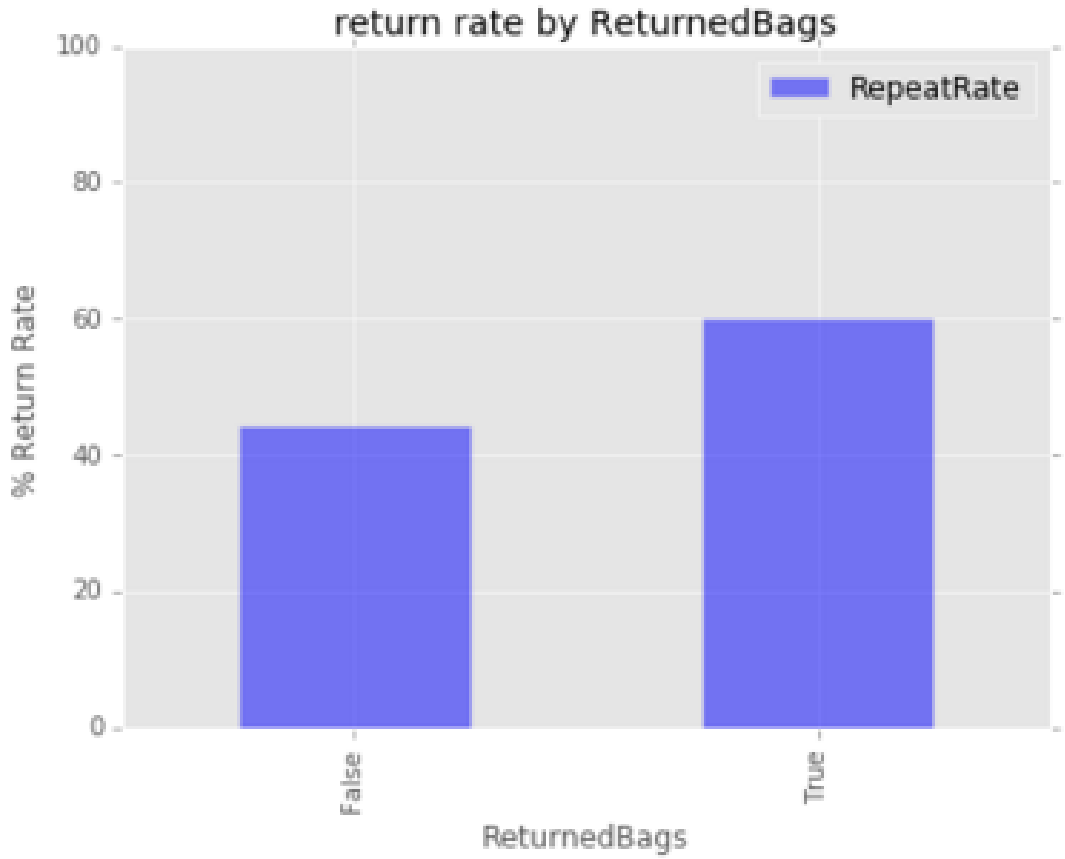


problem → EDA → models

repeataly

Some engineered features were not representative

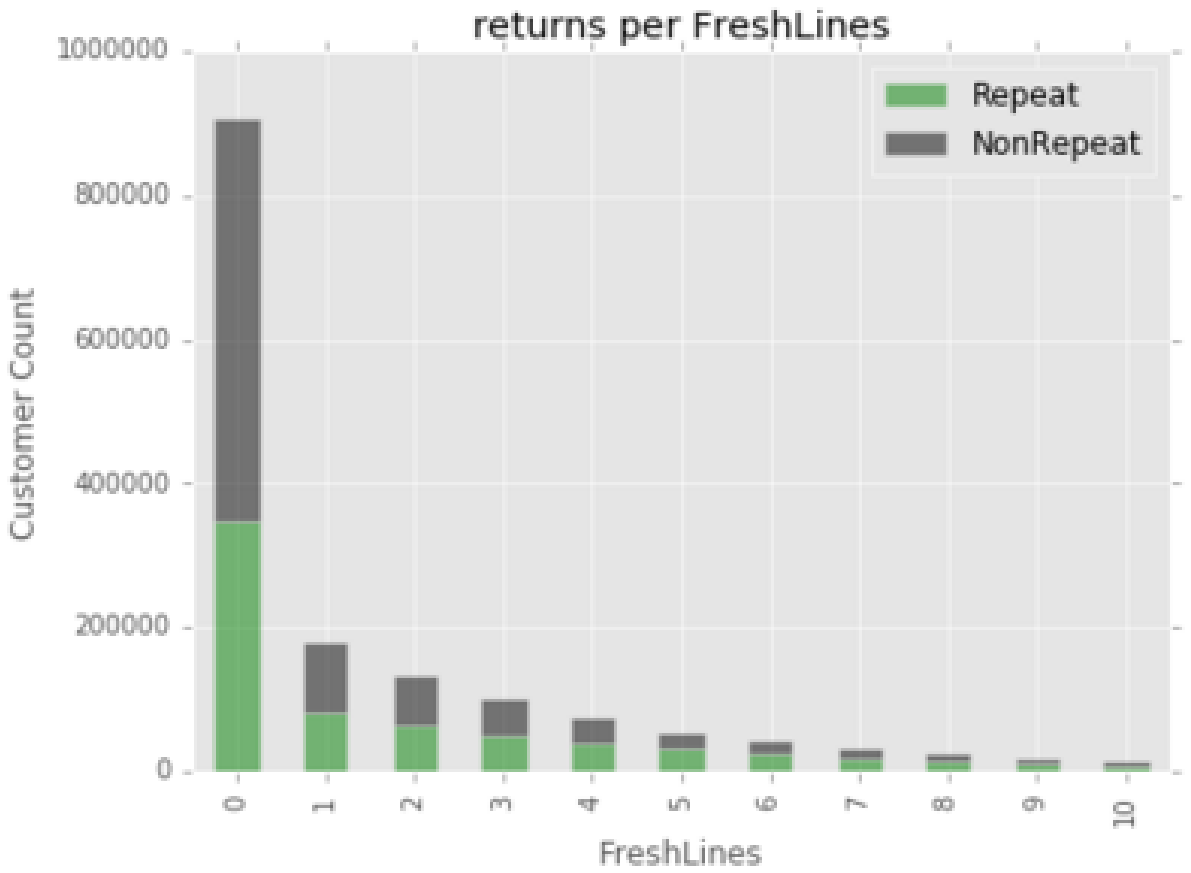


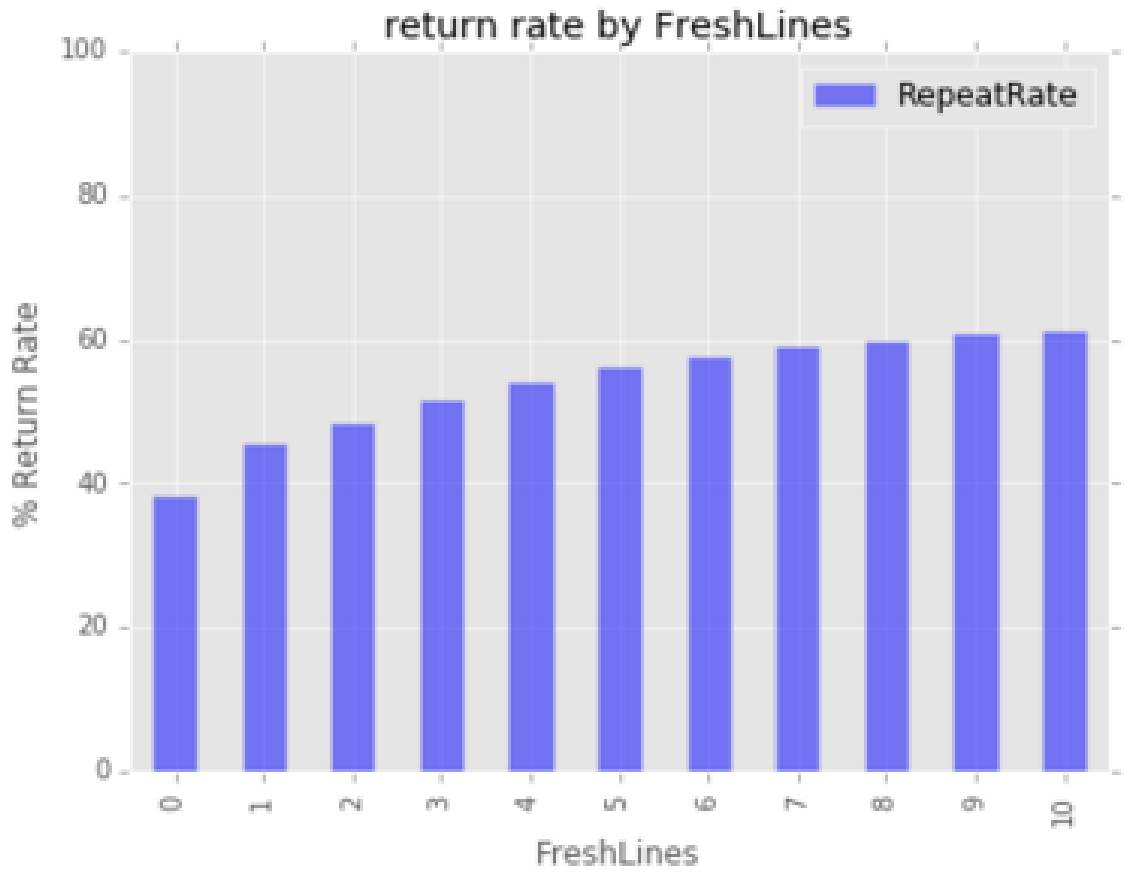


problem → EDA → models

repeataly

Fresh products seemed to correlate with repeat business

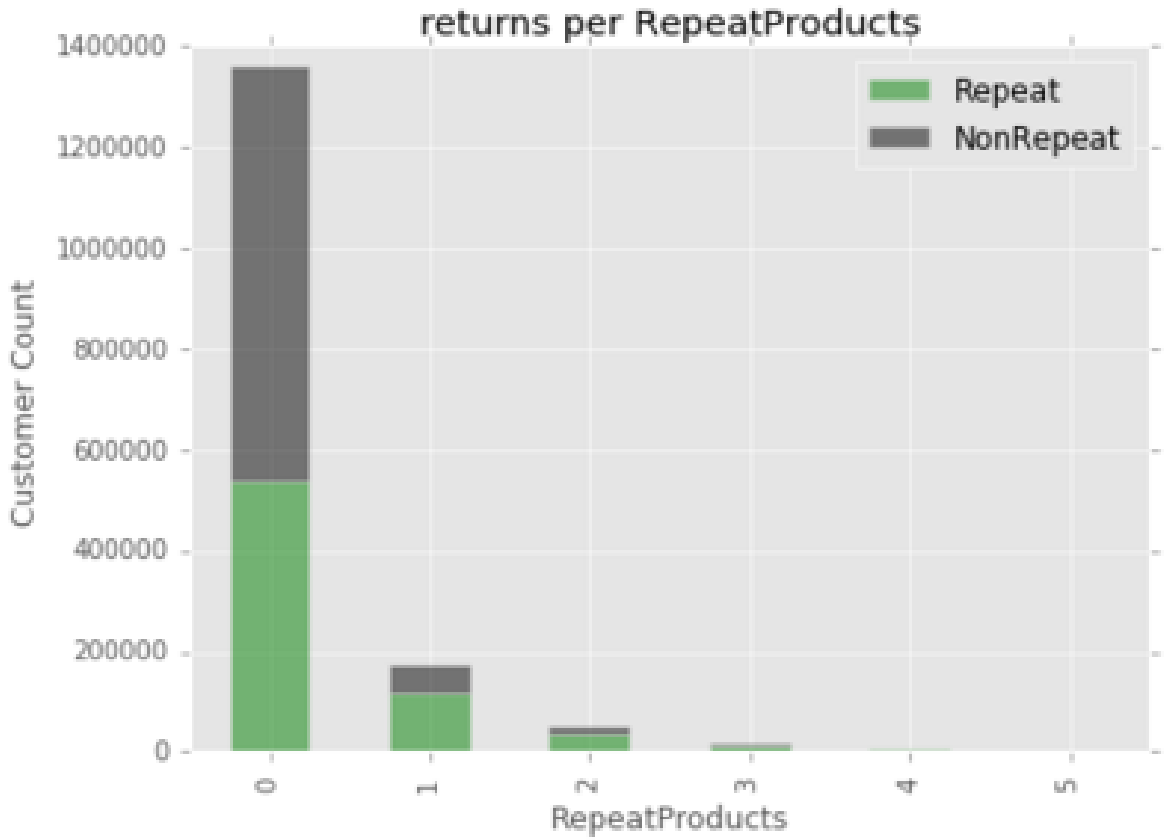


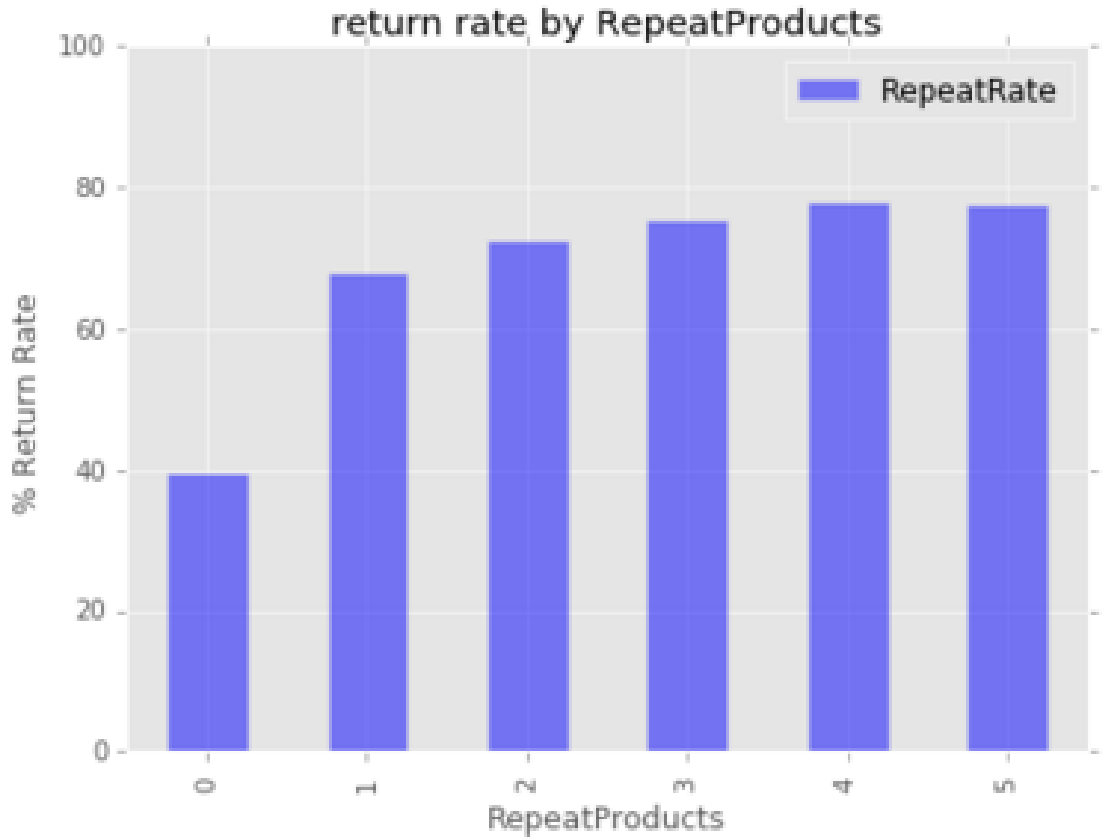


problem → EDA → models

repeataly

Repeat purchases suggest repeat visits!





problem → EDA → models

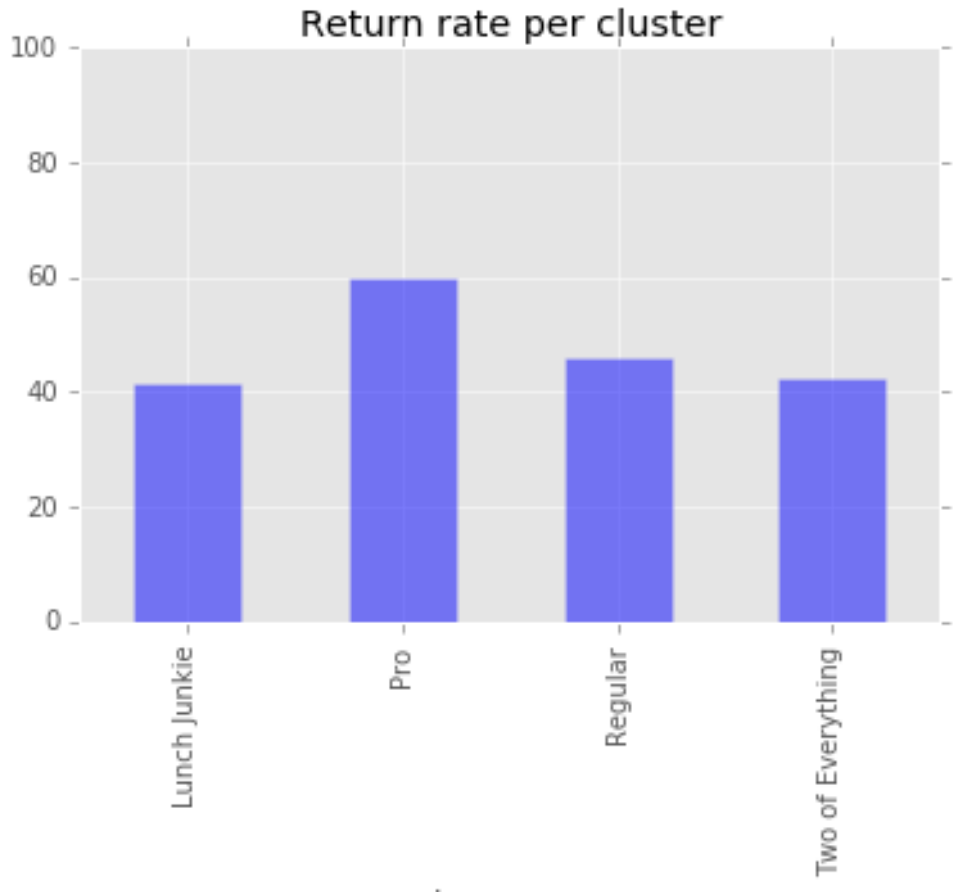
repeataly

Three models delivered more “scientific” feature selection

Dummy Classifier	Logistic Regression	Random Forest
Helped to set a baseline for predictive performance and feature importance.	Great measure of feature selection, and typically yielded the best performance in terms of speed and prediction	Also provided feature importance measures, and would probably be the production model as it handles correlated features better than LR

K-means was interesting but not a good feature

k-means clusters				
	0	1	2	3
Feature	**Two of everything**	**Regular**	**Lunch Junkie**	**Pro**
PriorVisits	6.536	6.454	3.908	7.502
NetAmount	38.653	34.342	12.97	157.862
TopItemLines	1.421	1.131	0.734	5.023
RepeatProducts	0.228	0.279	0.115	1.762
IsFrontEnd	0.719	0.995	0	0.998
FreshLines	1.612	2.598	0.166	12.187
RetailSpendCnt	0.967	0.948	0.127	4.328
ProductVariety	0.598	0.989	0.999	0.917
BizHours	0.384	0.328	0.447	0.364



Models scored very well

Dummy Classifier

Did not perform very well. Essentially taking the mean of the data set prediction.

	test_name	test_score
0	accuracy	50.480
1	precision	44.350
2	recall	44.350
3	f1	44.350

Logistic Regression

Best precision (**86%**), but model definitely needs exhaustive gridsearch on a more powerful machine.

	test_name	test_score
0	accuracy	70.220
1	precision	86.230
2	recall	39.350
3	f1	54.040

Random Forest

Great predictive performance across all metrics, but underlying features were clearly biased and need improvement

	test_name	test_score
1	precision	72.170
0	accuracy	72.100
3	f1	65.940
2	recall	60.710

LR and RF did not appreciate my features…

	coefficients	features
1	0.275	PriorVisits
7	0.014	RepeatProducts

Logistic Regression

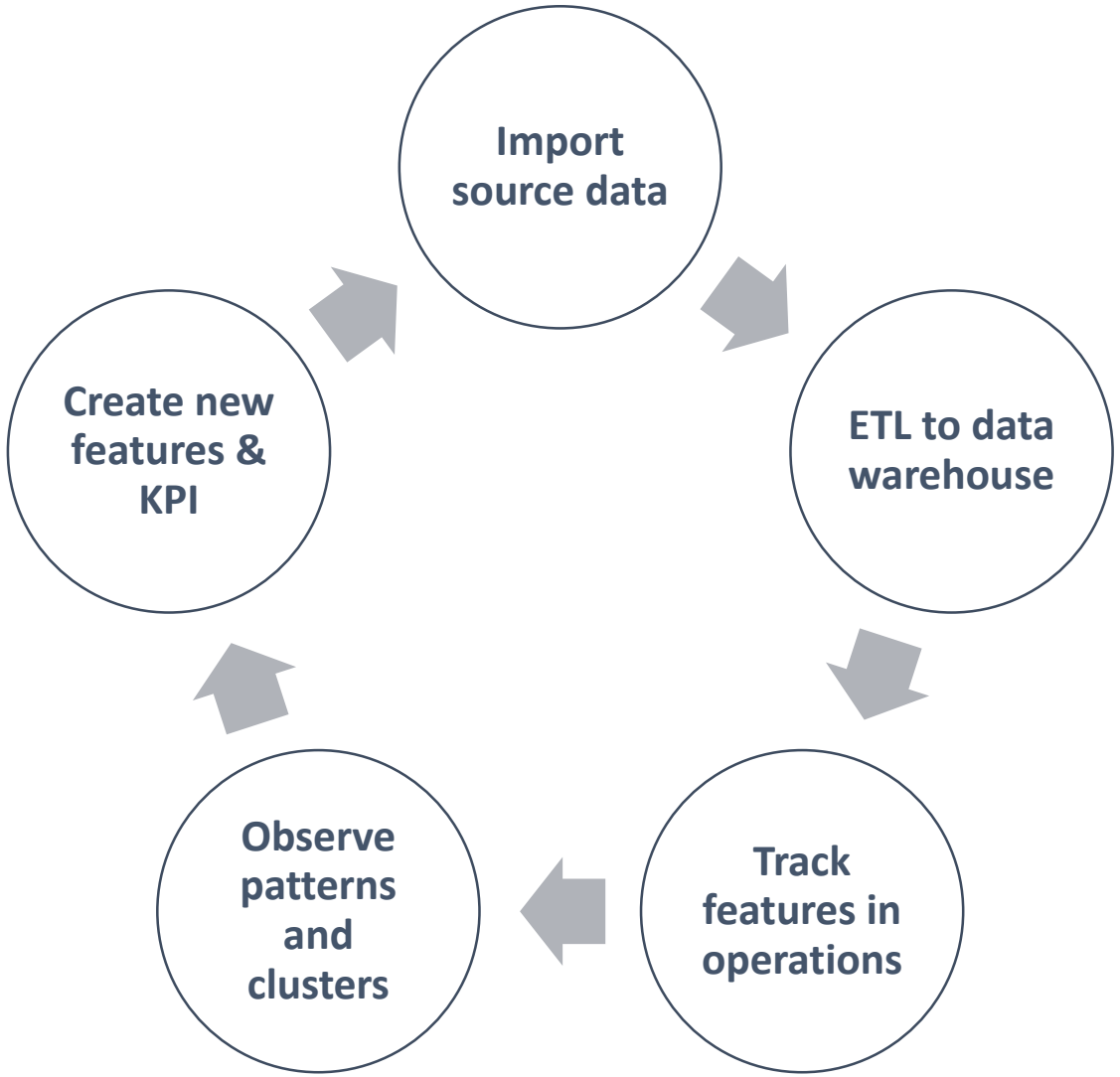
Needs new features, as highest coefficient was 0.275, and lasso only yielded two non-zero coefficients

	Features	Importance
1	PriorVisits	0.225
0	TicketTime	0.148
13	TicketCount	0.122
8	TicketMonthDecimal	0.118
2	NetAmount	0.109
12	RetailSpendCnt	0.084

Random Forest

Also needs new features. Prior visits outweighs the other features significantly, and no single feature is "actionable"

The goal is to track new features



project priorities and goals:

- **Interpretability of the models** – The research must yield actionable output. For this reason, random forest and logistic regression are employed alongside k-means clustering to segment customers and show feature importance.
- **Continuous implementation** – The long term goal is to understand features of customer behavior that drive sales and loyalty, and to track these features over time.
- **Find actionable features –** The other long term goal is to design features that operations teams can act on (as opposed to features designed for raw predictive power)

Questions?