Deliverables

1. Fast, automated churn prediction pipeline
2. Actionable, quantitative factors impacting churning
3. Recommendations for churn reduction

Profit Assumptions

Cost benefit matrix:

- Baseline is doing nothing
- Each user spends $50/month

	Actual		
Predicted		Churn	No Churn
	Churn	$30	-$10
	No Churn	0	0

Assumption:

- $10 coupon for predicted churners
- Retain 80% churners for +1 mo.

Purpose:

This allows us to maximize expected finance return!

Baseline “You know nothing, John Snow”

What if… we just send coupons to everyone??

(**62%** of users in the provided dataset have churned)

Accuracy	62%
Recall	100% - we correctly predict everyone who churns
Precision	62% - 62% of the people we predict churns are actually churn

<u>BUT that’s not the bottom line</u> Profit for user: **$14.8** - *We can do better*

Final Model

Logistic Regression: chosen based on highest maximum profit under initial assumptions.

Accuracy	75%
Recall	84%
Precision	78%

Profit per user: **$15.4** > baseline of $14.8

For our user-base of 40 million, our expected profit compared to sending coupons to all users is $24 million

Insights/Recommendations

Average Distance:

For every additional 10 miles, a user is 25% more likely to churn.

Trips in first 30 days:

For every additional trip in first 30 days, a user is 5% less likely to churn – give sign-up promotions! Make them get familiar with your app!

City:

Astapor are 68% more likely to churn than Winterfell and King’s landing are 71% less likely to churn than Winterfell

Insights/Recommendations

Phone:

Iphone users are 46% less likely to churn than unlisted phones. Android users are 38% more likely – target mobile ads toward iPhone

Weekday or Weekend:

Users who ride exclusively on weekdays or weekend are 4 times more likely to churn than those who use the service equally - discount for # of rides in a week

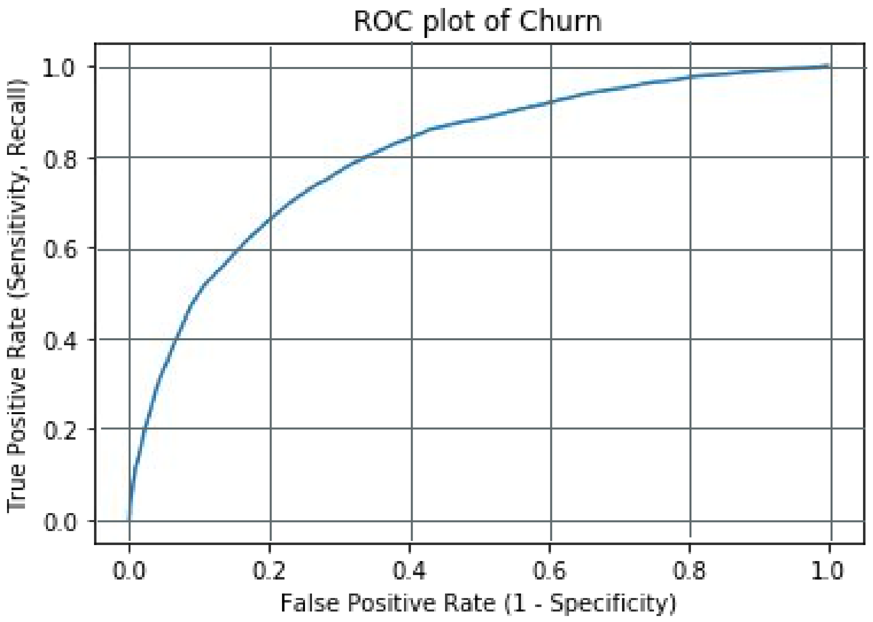
Luxury car users:

Luxury car users are 56% less likely to churn – discounted rides for users to experience luxury service

Final Model

Significant features:

Average distance:	0.0226	p-value = 0
# trip in first 30 days:	-0.0534	p-value = 0
City - Astapor:	0.5268	p-value = 0
City - King's Landing:	-1.2539	p-value = 0
Phone - iphone:	-0.7696	p-value = 0
Phone - Android:	0.3190	p-value = 0.024
Exclusive weekday user :	1.4004	p-value = 0
Exclusive weekend user:	1.59004	p-value = 0
luxury_car_user :	-0.8357	p-value = 0
Constant:	0.7181	



Performance on Test Data:

Accuracy:	0.7426
Recall:	0.8312
Precision:	0.7827

	Churn	No Churn
Churn	$30	-$10
No Churn	0	0

Threshold given our assumption = 0.3

Future Considerations

To make a better model, we would like some:

- Business data
  - Average profit, number of trips per customer
  - Traffic density
- Data of users
  - Trip date, distance, location, etc.
- Feedback data
  - Cost per user for coupons or promotion campaigns?
  - Did the churned user return?

Appendix

Results for first model

Logit Regression Results

Dep. Variable:	churn	**No. Observations:**	30000
Model:	Logit	**Df Residuals:**	29996
Method:	MLE	**Df Model:**	3
Date:	Fri, 21 Jul 2017	**Pseudo R-squ.:**	0.03988
Time:	11:54:20	**Log-Likelihood:**	-19070.
converged:	True	**LL-Null:**	-19862.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
avg_surge	0.0818	0.055	1.487	0.137	-0.026	0.190
avg_dist	0.0280	0.002	11.341	0.000	0.023	0.033
trips_in_first_30_days	-0.1335	0.004	-31.547	0.000	-0.142	-0.125
constant	0.5633	0.064	8.813	0.000	0.438	0.689

Results for second model:

Logit Regression Results

Dep. Variable:	churn	**No. Observations:**	30000
Model:	Logit	**Df Residuals:**	29995
Method:	MLE	**Df Model:**	4
Date:	Fri, 21 Jul 2017	**Pseudo R-squ.:**	0.09596
Time:	12:02:28	**Log-Likelihood:**	-17946.
converged:	True	**LL-Null:**	-19851.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
avg_dist	0.0321	0.003	12.715	0.000	0.027	0.037
trips_in_first_30_days	-0.1334	0.004	-30.654	0.000	-0.142	-0.125
city_Astapor	0.4169	0.030	13.967	0.000	0.358	0.475
city_King's Landing	-1.2111	0.033	-36.845	0.000	-1.276	-1.147
constant	0.7646	0.026	29.398	0.000	0.714	0.816

Appendix

Results for Third model

Logit Regression Results

Dep. Variable:	churn	**No. Observations:**	40000
Model:	Logit	**Df Residuals:**	39993
Method:	MLE	**Df Model:**	6
Date:	Fri, 21 Jul 2017	**Pseudo R-squ.:**	0.1335
Time:	14:21:40	**Log-Likelihood:**	-22943.
converged:	True	**LL-Null:**	-26479.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
avg_dist	0.0333	0.002	14.751	0.000	0.029	0.038
trips_in_first_30_days	-0.1325	0.004	-34.513	0.000	-0.140	-0.125
city_Astapor	0.4531	0.027	17.095	0.000	0.401	0.505
city_King's Landing	-1.1973	0.029	-40.904	0.000	-1.255	-1.140
phone_Android	0.5329	0.131	4.062	0.000	0.276	0.790
phone_iPhone	-0.5892	0.130	-4.544	0.000	-0.843	-0.335
constant	1.0288	0.131	7.869	0.000	0.773	1.285

Results for fourth model:

Logit Regression Results

Dep. Variable:	churn	**No. Observations:**	40000
Model:	Logit	**Df Residuals:**	39990
Method:	MLE	**Df Model:**	9
Date:	Fri, 21 Jul 2017	**Pseudo R-squ.:**	0.2235
Time:	14:15:13	**Log-Likelihood:**	-20562.
converged:	True	**LL-Null:**	-26479.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
avg_dist	0.0226	0.002	9.177	0.000	0.018	0.027
trips_in_first_30_days	-0.0534	0.004	-14.478	0.000	-0.061	-0.046
city_Astapor	0.5268	0.028	18.527	0.000	0.471	0.583
city_King's Landing	-1.2539	0.032	-39.627	0.000	-1.316	-1.192
phone_Android	0.3190	0.141	2.258	0.024	0.042	0.596
phone_iPhone	-0.7696	0.140	-5.508	0.000	-1.044	-0.496
Weekend	1.5969	0.037	42.614	0.000	1.523	1.670
Weekday	1.4004	0.029	48.634	0.000	1.344	1.457
luxury_car_user_True	-0.8357	0.025	-33.554	0.000	-0.885	-0.787
constant	0.7181	0.142	5.063	0.000	0.440	0.996

EDA insights

Phone type and number of trips in the first 30 days drastically alter retention rates

T = trips in first 30 days	Phone type count	Retention % (1-churn)	Retention % for T = 0	Retention % for T >= 3	Retention % for T >= 10
iPhone	27628	44.9%	42.8%	64.2%	78.9%
Android	12053	20.9%	21.4%	33.5%	51.3%
N/A	319	30.7%	43.3%	54.4	58.8%

One time users?

Whole Number for the ratings or NaN

Trips in the first 30 days <= 1

Weekday % is 0 or 100

Surge is 1.0 and surge % is 0

13026 users. (1594 if excluding 5.0 rating)

Only Weekday % is 0 or 100

20671 users

First Model

Logistic Regression for interpretability

predictors:

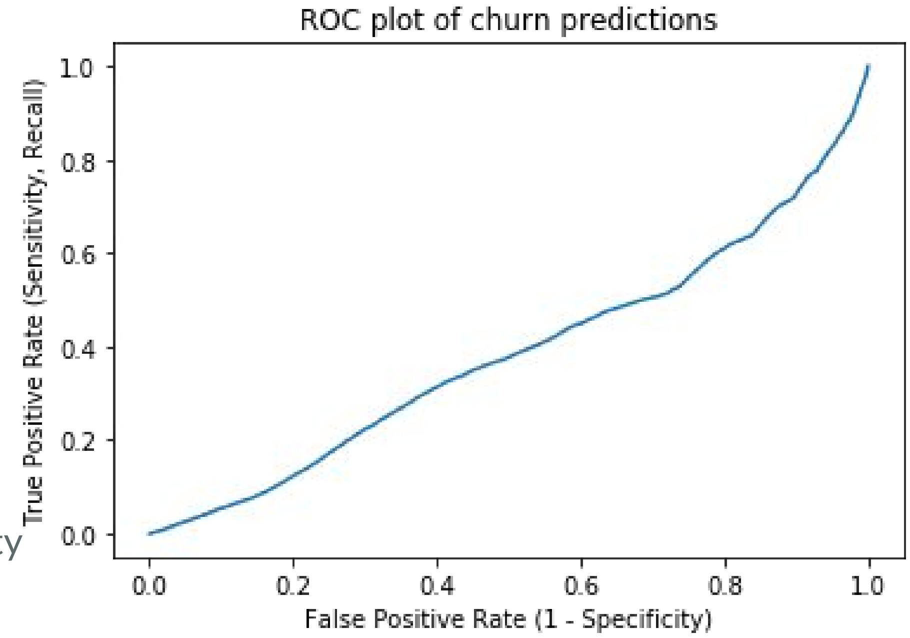
- Number of trips in the first 30 days
- average surge multiplier over all of this user’s trips
- average distance per trip taken in the first 30 days

First Model

Significant features:
Average distance: 0.0280 (p-value = 0)
trip in first 30 days: -0.1335 (p-value = 0)

Not significant:
Average surge: 0.0818 (p-value =0.137)

Accuracy: 0.656
recall: 0.948
precision: 0.655



Predicting 90% of the data points as 1’s - the majority class

Very Bad ROC curve...

Second Model

Logistic Regression for interpretability

predictors:

- Number of trips in the first 30 days
- average distance per trip taken in the first 30 days
- City

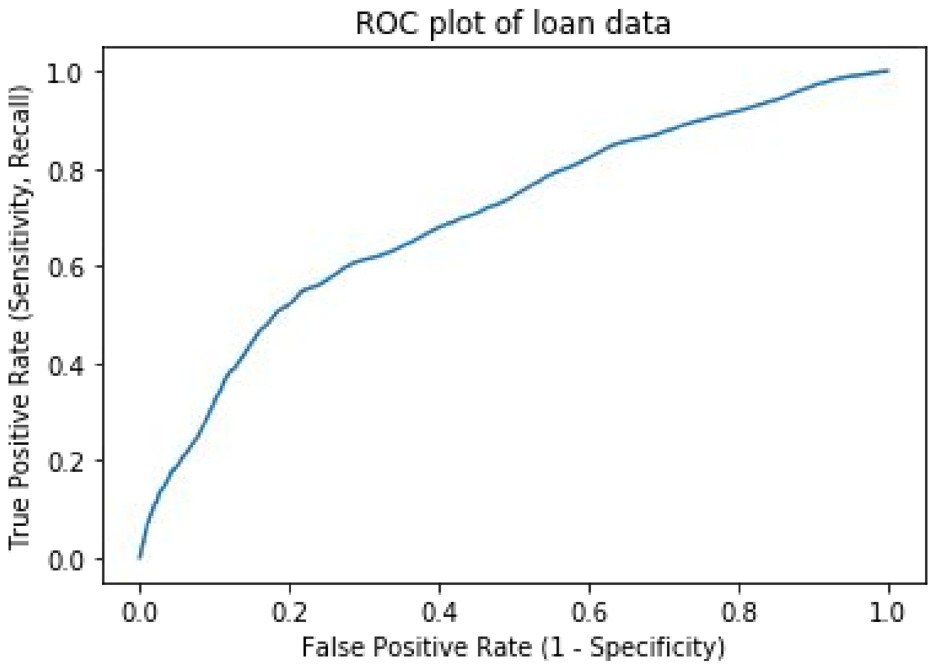
Second Model

Significant features: = ALL
Average distance: 0.0321 (p-value = 0)
trip in first 30 days: -0.1335 (p-value = 0)
City - Astapor: 0.4169 (p-value = 0)
City - King's Landing: -1.211 (p-value = 0)
Constant: 0.7646

Performance
Accuracy: 0.693
recall: 0.864
precision: 0.71

Predicting 76.06% of the data points as 1's - the majority class

Better ROC curve:



Third Model

predictors:

- Number of trips in the first 30 days
- average distance per trip taken in the first 30 days
- City
- Phone

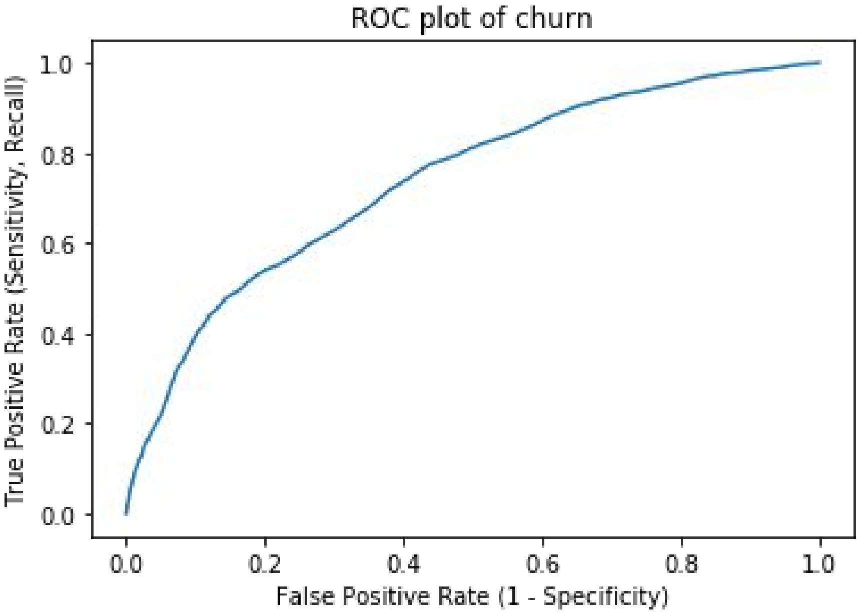
Third Model

Significant features: = ALL
Average distance: 0.0322 (p-value = 0)
trip in first 30 days: -0.1295 (p-value = 0)
City - Astapor: 0.4395 (p-value = 0)
City - King's Landing: -1.1883 (p-value = 0)
Phone - iphone: -0.6557 (p-value = 0)
Phone - Android: 0.4764 (p-value = 0.002)
Constant: 0.7646

Performance
Accuracy: 0.711
recall: 0.89
precision: 0.72

Predicting 77% of the data points as 1's - the majority class

Better ROC curve:



Random Forest Model

Best parameters: {untuned}

Best score: 0.663775

Best features: [0.61720939 0.27034737 0.11244324]

accuracy_score: 0.6657

prescision_score: 0.716104868914

recall_score: 0.767501605652

avg_dist , surge_pct, trips_in_first_30_days