Company X: Reducing Churn

Presented by Team Awesome:
John, Wall, Steve V Iannaccone,
Gretchen Riggs, & Brent Lemieux

Case Study

Company X has asked us to do some data analysis on their user data to help them understand what factors are the best predictors of retention:

- City?
- Phone type?
- Signup Date?
- Last Trip Date?
- Average Trip Distance?
- Average Rating by User
- Average Rating by Driver?
- % Trips when Surge Multplier > 1?
- Average Surge Multiplier over User's Trips?
- Number of Trips in first 30 days?
- Luxury Car User?
- % Trips during Weekdays?

1. Data Cleanup

- Changed any dates from String to Datetime type
- Changed City and Phone type to dummy variables (0, 1, 2, …)
- Changed NaN values to mean of feature
- After creating 'Active' feature (active users 06/01 /14-07/01/14, dropped date columns
- Response variable (y) = Churn user classification

What intuitively would make sense for having an effect on churn rate?

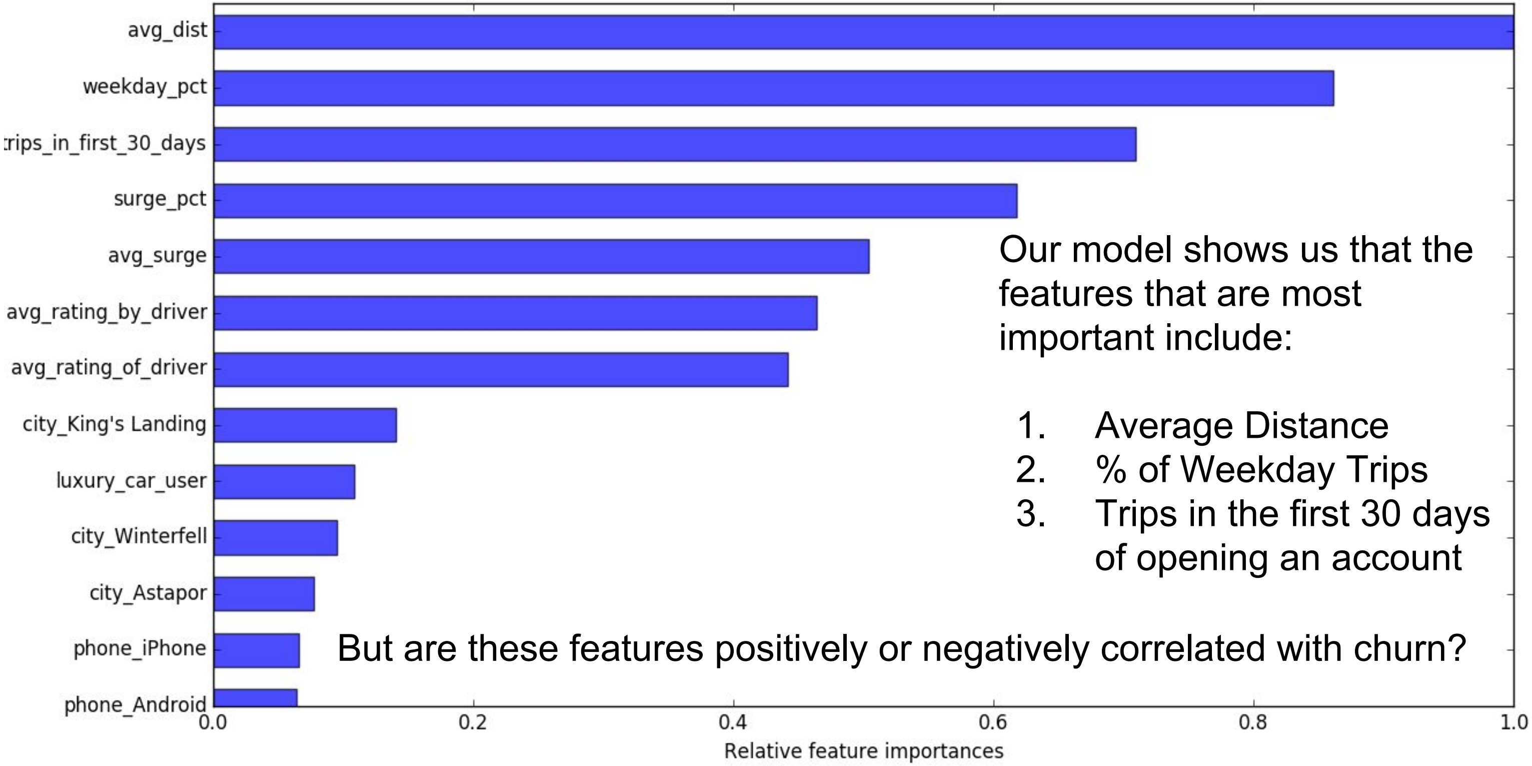
- Average Trip Distance
  - Short-trip users may use the service more due to low cost
- Average Rating by User
  - If users have a good experience with the service, more likely to stick around
- Percentage Trips when Surge Multiplier > 1
  - If users are paying too much for rides during times with a high surge rate, they may not stick around for more rides
- Number of Trips in First 30 days
  - Good first impression of service may lead to continued use
- Luxury Car User
  - These users are not as worried about cost of service, so may stay around even during surge times
- Percentage of Weekday Trips
  - If users regularly use service for commutes, they may stay around

- For this classification problem we explored using the following algorithms:
  - Random Forest
  - K Nearest Neighbors
  - SVC
  - Logistic Regression
  - Neural Network
  - AdaBoost
  - Gradient Boost

This was the winner!

It produced the highest F1-Score of all our test models of 0.84.

Feature Importance



Our model shows us that the features that are most important include:

1. Average Distance
2. % of Weekday Trips
3. Trips in the first 30 days of opening an account

But are these features positively or negatively correlated with churn?

Logit Regression Results

==

Dep. Variable:	y	No. Observations:	30000
Model:	Logit	Df Residuals:	29987
Method:	MLE	Df Model:	12
Date:	Fri, 20 Jan 2017	Pseudo R-squ.:	0.1602
Time:	16:09:15	Log-Likelihood:	-16702.
converged:	True	LL-Null:	-19888.
		LLR p-value:	0.000

==

	coef	std err	z	P>|z|	[95.0% Conf. Int.]	
const	0.5908	0.014	43.259	0.000	0.564	0.618
x1	0.2068	0.015	13.439	0.000	0.177	0.237
x2	0.0818	0.013	6.138	0.000	0.056	0.108
x3	0.0296	0.013	2.247	0.025	0.004	0.055
x4	0.0355	0.023	1.542	0.123	-0.010	0.081
x5	-0.0710	0.022	-3.165	0.002	-0.115	-0.027
x6	-0.4766	0.017	-27.422	0.000	-0.511	-0.443
x7	-0.4203	0.013	-31.439	0.000	-0.446	-0.394
x8	0.0080	0.014	0.582	0.561	-0.019	0.035
x9	0.3054	4.4e+05	6.94e-07	1.000	-8.62e+05	8.62e+05
x10	-0.4356	3.76e+05	-1.16e-06	1.000	-7.38e+05	7.38e+05
x11	0.0636	4.66e+05	1.36e-07	1.000	-9.14e+05	9.14e+05
x12	0.1117	0.072	1.553	0.120	-0.029	0.253
x13	-0.3922	0.072	-5.472	0.000	-0.533	-0.252

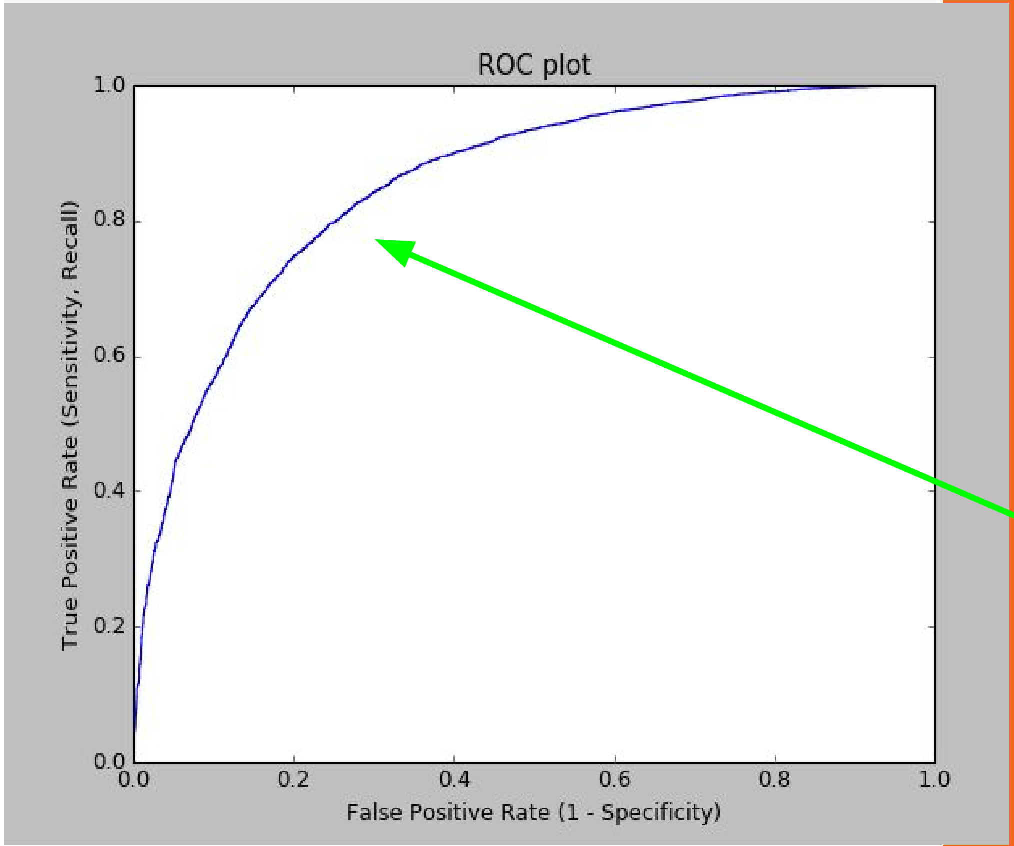
x1 = Average Distance
=> pos corr

x8 = % of Weekday Trips
=> pos corr

x6 = Trips in the first 30 days of opening an account
=> neg corr

By looking at the coefficients for each feature of our preliminary results using Logistic Regression, we can determine if the features are correlated positively or negatively with the output, churn. We are just looking at positive & negative relationships here.

ROC Curve Analysis

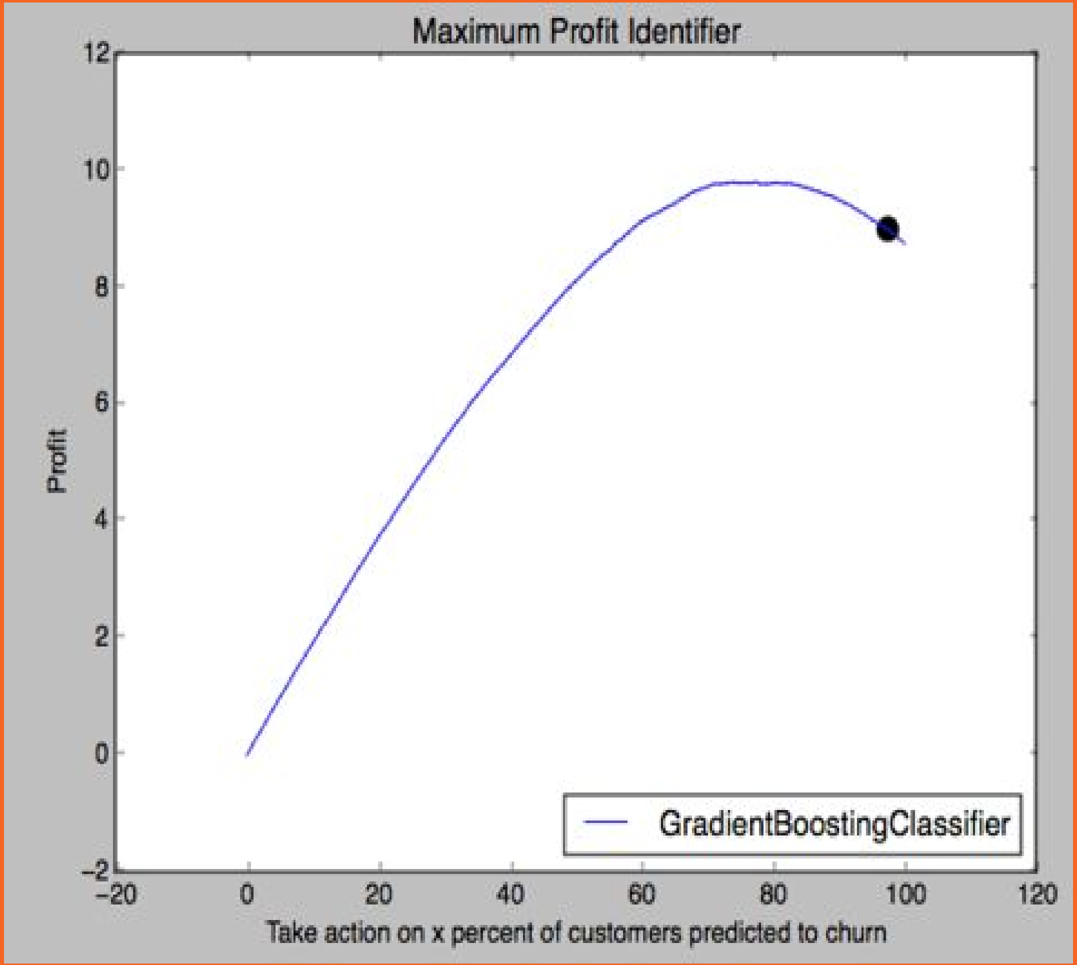


Actual

Predicted	Churn	Not Churn
Churn	TP = 5406	FP = 1257
Not Churn	FN = 843	TN = 2494

Accuracy = 0.79
Precision = 0.81
Recall = 0.86
F1-Score = 0.84

From the ROC curve, computed on our test data, which was not used in training our model, our predictions show a low level of False Positives and a high level of True Positives.

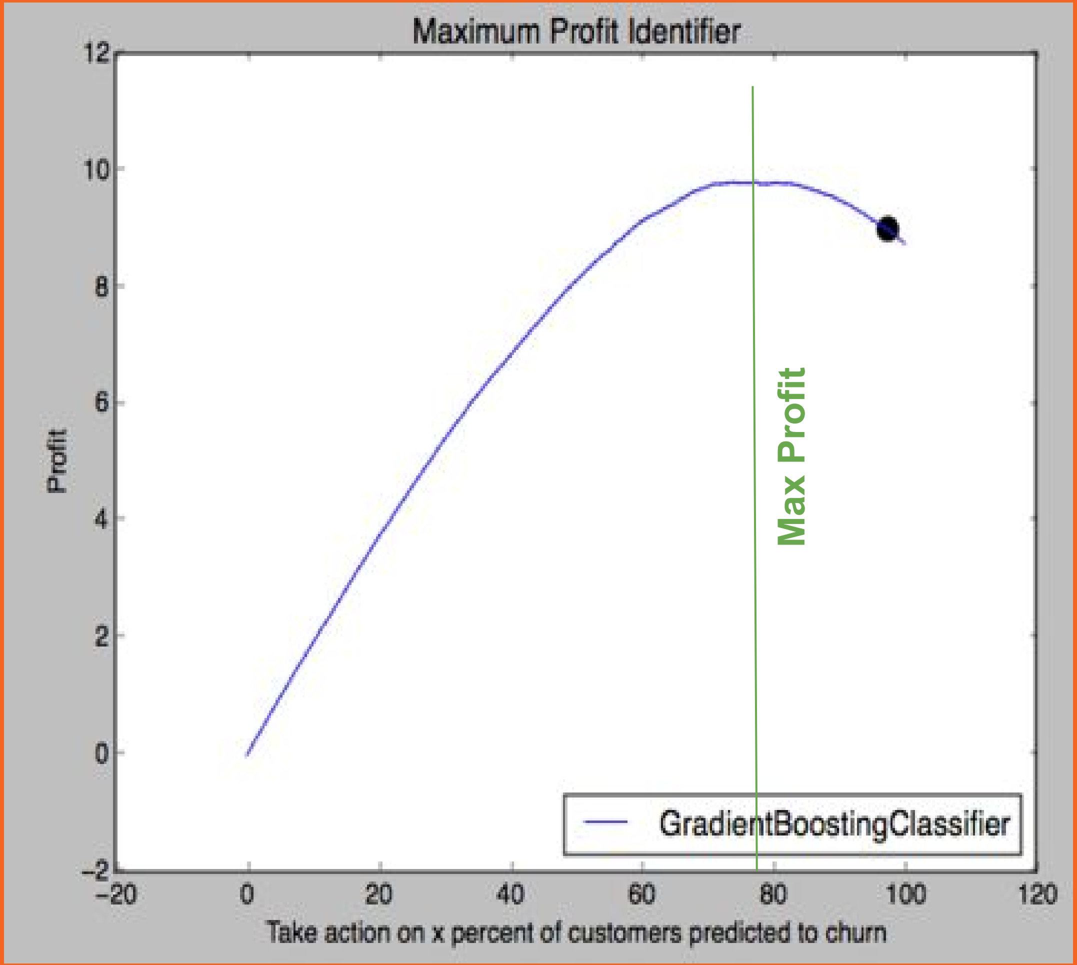


Cost Benefit Matrix
Actual

		Churn	Not Churn
Predicted	Churn	$20	-$10
	Not Churn	$0	$0

Assumption for this Case Study:
If we correctly predict customer who is going to churn and proactively send them a discount card, costing us $10, we profit an additional $20.

If we incorrectly predict customer will churn and send them a discount card and they don’t churn, we lose $10 in marketing costs.



Cost Benefit Matrix

Actual

Predicted		Churn	Not Churn
	Churn	$20	-$10
	Not Churn	$0	$0

To maximize profits, we can take action by sending ~75% of the users we predicted to churn the discount cards, if it is within the company's budget.

Gradient Boost code in Python

```
def gradientboost(X, y):
    X_train, X_test, y_train, y_test = train_test_split(X, y)
    est = GradientBoostingClassifier(n_estimators=1000, learning_rate= 0.1, max_depth=3, max_features='sqrt')
    model = est.fit(X_train,y_train)
    pred = model.predict(X_test)
    acc = model.score(X_test, y_test)
    sort_feat = np.argsort(model.feature_importances_)[::-1]
    report = classification_report(y_test, pred)
    matrix = conf_mat(pred, y_test)
    print 'acc', acc
    TP, FN, FP, TN = matrix[0][0], matrix[0][1], matrix[1][0], matrix[1][1]
    recall = float(TP)/(TP+FP)
    precision = float(TP)/(TP+FN)
    print 'precision', precision
    print 'recall', recall
    print 'f1', 2*(precision*recall)/(precision+recall)
    return model
    # params = {'n_estimators': [100, 500, 1000],
    #               'learning_rate': [.01, .1, .5, 1], 'max_depth': [1, 2, 3],
    #               'max_features': ["sqrt"]}
    #
    # grid = GridSearchCV(est, params, scoring='f1')
    # gridfinal = grid.fit(X_train, y_train)
    # return gridfinal
```

Conclusions:

- The following features are the highest correlated indicators of churn:
  - Average Trip Distance

    The longer the average trip, the more likely a customer is to churn, which could be because these users are only using the service for the occasional trip to the airport or similar.

  - % of Weekday Trips

    If a customer has more weekday trips, they may have just been using the service more occasionally, based on the numbers. The weekend customers must be more of our user base.

  - Trips in the first 30 Days

    This feature is negatively correlated with churn, so if a customer has many trips in the first 30 days, they are less likely to churn.

Actions:

- Using our predictions, we can target our marketing campaign to the users we predict to churn.
- We can maximize our profits by targeting 75% of these users whom we predict will churn.