Model evaluation

Our model evaluation process entails several important steps in order to tackle the complexity of our data set. The complexities mainly arise from:

- Highly imbalanced data set
- Mixture of numeric and categorical features
- Self-engineered Y labeling

To tackle these challenges we employed two versions of labeling technique (version 1 and version 2) and conducted a number of cross validation tests on up-sampling values to evaluate which method is most suitable for our data set. [recap of what each version does]

Model Selection

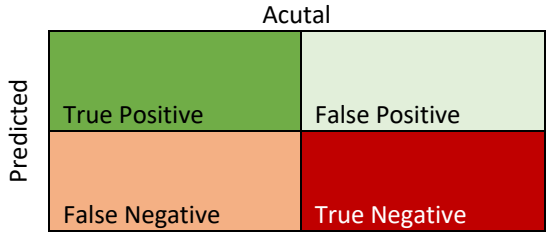
For model selection we considered two different types of models: Logistic Regression and Decision Tree. We felt that Logistics Regression might be suitable because we have a two-class classification and our number of data points (n) is much greater than our number of features (m). If fact, our n = 6,438,528 and our m= 89

We also felt that Decision Tree could work well because we felt that the features were not independent of each other, and that Decision Tree would be able to consider the relationship across features to determine an instance's class. For example, a station that is located in a neighborhood that is busy all the time, then dayOfWeek (whether it is a weekday or not) may not matter very much, while a station located in Midtown East, where many people work there, dayOfWeek would be a more important feature in categorizing the station is at risk or not.

To systematically decide which model would be more suitable for our project, we decided to use 3 random subsets of Station IDs as model selection test cases. We used 50 stations first as opposed to our overall data set has 741 Station IDs and this would be a quick way to check which model works well before going deeper into cross validation and finetuning our model parameters on the broader data set.

Because we have a very unbalanced classification problem, we used the confusion matrix as opposed to accuracy score to decide which model works better. It became very clear that Decision Tree would be the better model because Logistic Regression was not able to identify any class=1 instances.

Confusion Matrix Legend



Comparison of Logistic Regression and Decision Tree Performance on Subsets of 50 Station IDs

	Logistic Regression		Decision Tree	
first 50 station ID set	0	0	2	263
	222	88098	220	87835
second 50 station ID set	0	0	6	217
	204	88116	211	87899
third 50 station ID set	0	0	9	202
	204	88116	195	87914

We think Logistic Regression did poorly because we have a number of categorical features that may be difficult for Logistic Regression to model. Further we think the fact that our data is so imbalanced may also pose a challenge for Logistic Regression.

From this initial analysis, we decided to deploy our version 2 technique in Y-labeling. [short explanation of version 2]. The main benefit of using version 2 versus version 1 is that the percentage of positive classes in version 1 is 0.24%, where that percentage is 0.68% for version 2. In addition, version 2 may be a 'strong' at risk flag because it flags a period of continuous 'at risk' time buckets as opposed to the first time bucket of a 'at risk' period. Refer chart [we should refer to some chart here to describe the difference]

Cross Validation

The main objectives of our Cross Validation are to decide which version of our Y-labeling technique provided better results and also which level of up-sampling is most sensible. Below we offer side-by-side comparison of the two version of Y-label techniques. It is clear that version 2 of Y-labeling offers a more balanced outcome between Recall and Precision.

Version 1	Recall	Precision	F1 score	FPR
no up sampling	0.04	0.03	0.03	0.00
1% positive	0.11	0.03	0.04	0.01
2% positive	0.19	0.02	0.04	0.02
5% positive	0.32	0.02	0.04	0.04
10% positive	0.44	0.02	0.03	0.07
20% positive	0.58	0.01	0.02	0.11

Version 2	Recall	Precision	F1 score	FPR
no up sampling	0.21	0.39	0.28	0.00
1% positive	0.28	0.35	0.31	0.00
2% positive	0.44	0.28	0.34	0.01
5% positive	0.64	0.22	0.33	0.02
10% positive	0.76	0.18	0.29	0.02
20% positive	0.86	0.14	0.24	0.04

Furthermore, when deciding which up-sample level to use.

We are dealing with a highly unbalanced set of data in both Version 1 and Version 2 of our data set. Hence checking the accuracy of the models alone with not be sufficient in deciding which model gave a better prediction, because any prediction that assigns a large number of negative class would receive a high accuracy score.

As such we feel that comparing the tradeoff between recall and precision is very important in our case. In our evaluation below, we calculated the recall and precision of each model. We also provide the F1 score here, which is the harmonic mean of precision and recall. However, we want to show the underlying recall and precision tradeoffs because management can couple that with their information on the cost and benefit of fleet management to make the best business decision.

Selection of our Labelinhema

- Effects on

Given that our model has very few actual positive instances
Version 2: 070%
Version 1: 0.24%

Effects of Up Sampling results
Compare side by side of version 1 and version 2

We seek to find a model that has a higher true positive rate

There is a trade off between true positive rate and recall = tp/(tp + fn) the number of actual Y=1 that were predicted as 1s

Base rate:
The Base rate for our project is not straight forward because we engineered many features to track the traffic of each station and we would assume that Citibike management would use similar strategies to identify stations that have prolonged strong demand.

If we were to assume that Management also identified a time bucket as ‘at risk’ the same way that we do, then the base rate would be the total number of times the time bucket was flagged as ‘at risk’ across all the time buckets for 93 days.

Furthermore, management can focus stations that have very high activity i.e. variance is greater than 60% of the capacity at that station. Looking at those stations, we can check the ratio of times that each timebucket was flagged as ‘at risk’ and that would be a base rate.

<Insert table>

<Insert chart>