# Machine Learning:

## The Key to Outsmarting the Competition

MIT Sloan's Global Entrepreneurship Lab
Fall 2017, IAP 2018

### "Brazil's First Unicorn"

Sao Paulo, Brazil


C. Bull | N. Mehta | T. Shane | C. Wagner

## The Approach

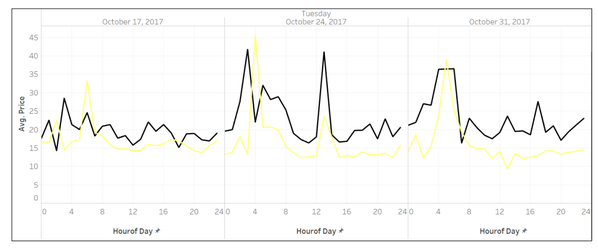
1. Evaluate the microeconomics of the e-hailing industry
2. Benchmark pricing theory for two-sided platforms and study the mechanics of network development
3. Leverage external data to create dependent variable
4. Utilize machine learning techniques to identify relationships with internal and external data sources
5. Generate dashboards to enhance data-driven decisions

## The Results

**Model**

```
#PREDICT ON TEST SET AND GET ACCURACY
Predicted_SVM <- predict(SVM, newdata = Test)

SVMMatrix <- confusionMatrix(Predicted_SVM, Test$Surge.Binary)
SVMAccuracy <- SVMMatrix$overall[["Accuracy"]]
SVMAccuracy

#CALCULATE IMPROVEMENT OVER BASELINE
improvement_svm = (SVMAccuracy-baseline_accur)/baseline_accur
improvement_svm

#AUC
Predicted_SVM_Num <- as.numeric(Predicted_SVM) #ROCR DOESN'T USE FACTORS
rocr.pred.svm <- ROCR::prediction(Predicted_SVM_Num, Test$Surge.Binary)
plot(ROCR::performance(rocr.pred.svm, "tpr", "fpr"), colorize=TRUE)
abline(0,1)

as.numeric(ROCR::performance(rocr.pred.svm, "auc")@y.values)

###BOOSTED TREES###
set.seed(123)

trainTask = makeClassifTask(data=Training,target='Surge.Binary')
testTask = makeClassifTask(data=Test,target='Surge.Binary')

getParamSet('classif.xgboost')

#CREATE BINARY CLASSIFICATION LEARNER AND SET INITIAL PARAMETERS
xg_set = makeLearner('classif.xgboost',
                     predict.type = 'prob')
```

**Dashboard**




| | |
|---|---|
| **17** Machine learning models utilized to maximize predictive power | **13** Interviews with members of Ops, Data Science, and Strategy teams |
| **77** Distinct features tested for relationship with dependent var. | **9** Dynamic visualizations filterable by date, time, city, and district |
| **5K** Observations sampled to create the dependent var. | **4M** Data points from e-hailing market |

**Off-Site**

### 1 Evaluate the microeconomics of the e-hailing industry

**High Substitutability:** 99 competes in a crowded urban mobility market where eHailing is still relatively small but growing quickly

**Little Product Differentiation:** Minimal product differentiation has led 99 and its competitors to focus on competing on price and availability to win market share

**Shifting Barriers to Entry:** Low initial barriers to entry create fierce short-run competition but high barriers are likely to arise in the long-run

In the **short-term**, these factors indicate an **oligopoly market** where firms have little market power and thus **compete primarily on price**

In the **long-term**, firms hope to outlast competitors and move to a **monopolistic market**, where they can generate significant profits with **higher prices**

### 2 Benchmark pricing theory for two-sided platforms and study the mechanics of network development

**Traditional Two-Sided Platform (e.g. Credit Cards)**

| Subsidy Side | | Money Side |
|---|---|---|
| To incentivize cardholder use, a negative fee is applied | $$ ← VISA ← $$ | Merchants pay a larger amount to offset the negative fee paid by users |

**E-Hailing Two-Sided Platform – A Case of Double Subsidies**

| Subsidy Side | | Subsidy Side |
|---|---|---|
| Riders are charged a rate at which P < MC in order to incentivize use | $$ ← UBER → $$ | Drivers are paid more for each ride than riders pay in order to build supply side of the network |

Pricing in this way (with two subsidy sides) is wholly contingent upon lavish rounds of external venture capital funding, and unsustainable in the long run.

### 3 Leverage external data to create dependent variable

| Step | Objective | Methodology |
|---|---|---|
| 1 | Eliminate Noise | All entries with data feed errors and/or incomplete and uninterpretable information were eliminated (~3M obs remained) |
| 2 | Reduce Variability in the Data | Only usage data for rides in Sao Paulo with no promos were utilized in order to limit presence of confounding variables (~80K obs) |
| 3 | Create Sample Data Set | A random sample of 5K observations were utilized in order to gauge viability of calculation methodology |
| 4 | Calculate Ride Duration and Distance | Data set was processed through Google Maps API in order to generate duration and distance estimates from start and stop points |
| 5 | Calculate Expected Fare | Given ride duration and distance, an expected ride fare was calculated with specified formula |
| 6 | Calculate and Process Dependent variable | Dependent variable was calculated by dividing observed fare by expected fare and then floored |
| 7 | Regress Dependent variable On Features | Dependent variable was regressed on a litany of features |

**On-Site**

### 4 Utilize machine learning techniques to identify relationships with internal and external data sources

Started with **77 possible features** to establish predictive relationships → Fed into → **17 Machine Learning Models**, spanning regression, CART, Random Forest, Nearest Neighbors, Support Vector Machines, and Boosted Trees

**4** categories of key variables: Supply, Demand, Temporal, Market Shocks → Yielded → Which resulted in → **~70%** Predictive Accuracy; **~20%** Improvement Over Baseline

### 5 Generate 3 dashboards to enhance data-driven decisions

**Historic Prices Dashboard**
- Week over week price comparisons
- Period determination (e.g. one week, two weeks, etc.)
- Product-selection

**Promotion Impact Dashboard**
- Week over week promotion impact comparisons
- Period determination (e.g. one week, two weeks, et al.)
- Product-selection

**Marketplace Shifts Dashboard**
- Insight into greatest pricing discrepancies between two market product offerings
- Time of day determination (e.g. morning rush, entire day, et al.)

### Next Steps — Ensure success through three key actions

| | | | |
|---|---|---|---|
| 1 | **Use the Dashboard** | GMs<br>Ops Teams<br>Data Science | The dashboard will be **available for use soon**; this is a fantastic **test and learn opportunity** to improve it moving forward |
| 2 | **Allocate Data Science Resources** | Engineering<br>Operations<br>Data Science | A small team from engineering, ops, and Data Sci. will be needed to launch; in steady state, **one analyst from ops and Data Sci. (ea.)** |
| 3 | **Communicate Success** | All 99ers! | Once operational, c**ommunicate new data driven approach widely**. Highlight wins/uses in GM Slack/Whatsapp groups |