GALANCK CASE:

HOW DATA CAN HELP BUSINESS

Final project for Data Analytics course presented by André Vitta

GALANCK

Connected backpack that combines GPS navigation, safety, ergonomy and style

MOBILITY - WITHOUT FEAR - WITHOUT RISK - WITHOUT POLLUTING

Who says we have to choose between style and safety? Who said good taste can't save lives?

HOW IT WORKS ?

FONCTIONNALITES CONNECTEES

FONCTIONNALITES INTEGREES

BIKE

TROTTINETTE

WALKING

PROJECT GOALS

IRON HACK

01 **PRODUCT DEVELOPMENT**

Improve Customer Product experience

02 **OPERATIONAL REQUIREMENTS**

User Data online processing and Storage

03 **TECHNICAL IMPROVEMENTS**

AutoNav routing

04 **FINAL REMARKS**

Make insightful propositions based on User data

OVERALL DIAGNOSTIC

I.6 % USERS

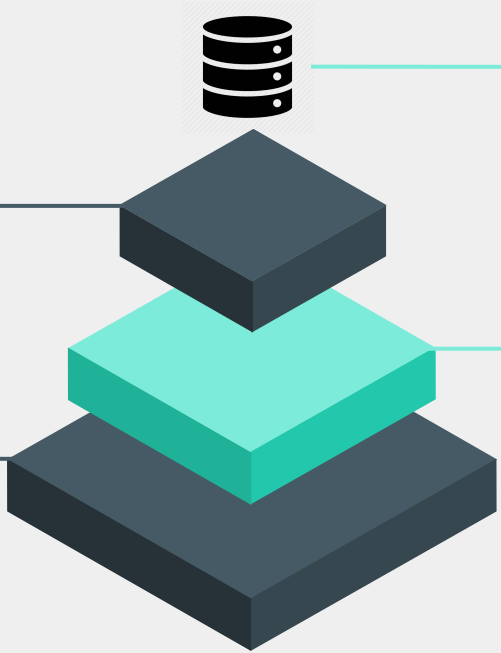
NAVIGATION AND PRODUCT DATA

Legal and Storage Issues

TOTAL USERS

PRODUCTS SOLD

Product just started at amazon



DATA STORAGE

BACKEND USER DATA

Storing data can be expensive especially if there is no use for it

35 % USERS

AUTO LIGHT AND NAVIGATION

Two thirds of customers not fully experiencing the product utility

QUESTIONS ?

- What is the “Value” of Data (if there is enough)
- What is the Reroute event rate?
- What are the reroute events situation ?
- What are the relevant Navigation statistics?
- How much improvement in operation can be achieved ?

HOW CAN DATA ANALYSIS HELP ?

- Understand the problem
- Get data
- Clean data
- Exploratory Data Analysis
- Feature Engineering
- Define and Run models
- Model Validation

UNDERSTAND THE PROBLEM

PRODUCT PERFORMANCE

Better routing algorithm can increase number of AutoNav users and their product perception

ADDRESS PROFESSIONAL USERS

High number of trips per day
Navigation precision has impact on productivity

DATA QUALITY PARAMETERS

Having this parameters can help to monitor product performance and development requirements

OPERATIONAL

Improving Data Storage and Estimating Cloud processing requirements

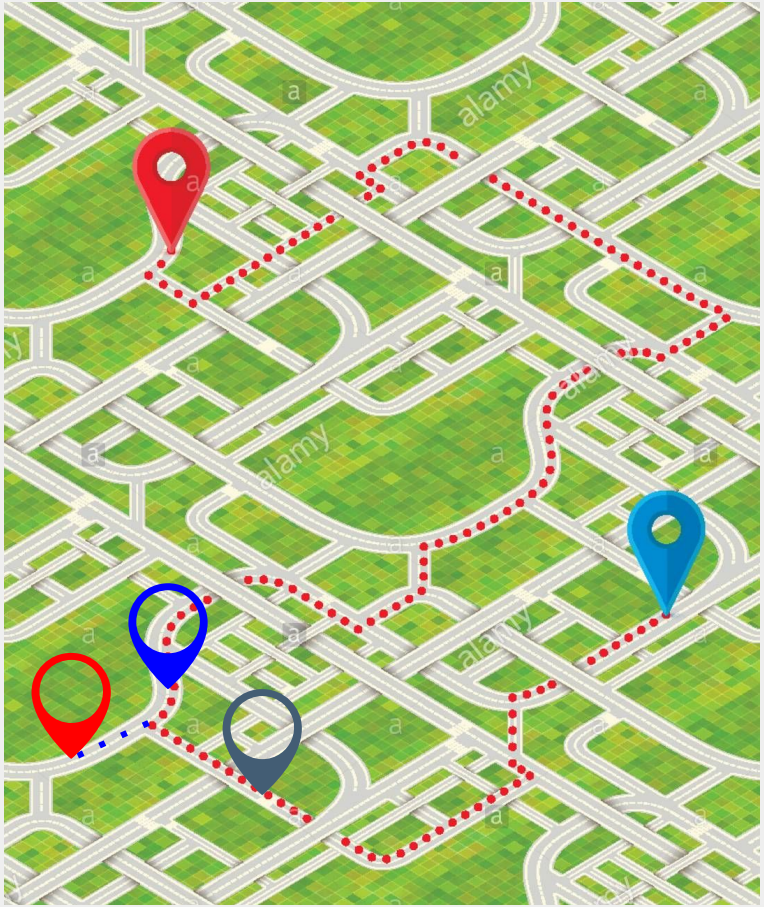
DATASET

TRIP DATA

- Plan route
- Actual route
- BLE Orders

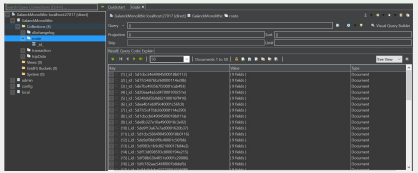
TERMINAL AND BAG INFO

- Bag ID
- Firmware version
- Phone model
- App Version
- App status
- Phone Baterry



HOW DATA IS COLLECTED

GPS DATA



BACKEND

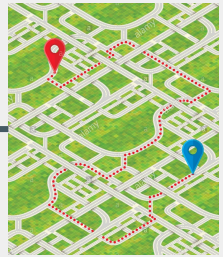
MONGODB

APP ON PHONE

BAG
DEVICE

FRONTEND

EVENT



Planned Route

DATA CLEANING AND MANIPULATION

327 TRIPS

Extracting GPS data and Events descriptions and Timestamps

TURNS CLASSIFICATION

Find all turns and verify if they were successful or not

CALCULATING ADDITIONAL PARAMETERS

Calculating some KPIs and some new variables as Gps Accuracy Variation and Time Gap

FINAL DATA

7 categorical features
10 numerical
1 target
2216 row and 18 features

GOOD TURN	**1323**
REROUTE	**893**

FEATURE ENGINEERING

FEATURES WITH DIFFERENT SCALES

Lat, Lon - Degrees
Distance - Meters
Time Gap - ms
Categorical Data - 1 to 5
Target - 0 and 1

NO PCA !!!

Since the goal is to understand how features impact on rerouting, having PCA don't really help

DUMMIES

Users Event Types and OSVersion had multiple categories so they were dummy

MODEL DEFINITION

LoGit
KNeighborsClassifier
GaussianNB
DecisionTreeClassifier
RandomForestClassifier
XGBClassifier
SVC
LogisticRegression
CatBoostClassifier
NuSVC

RESULTS

	Accuracy Score	F1 Score	Recall Score	Precision Score
GaussianNB	0.45	0.59	0.98	0.42
KNeighborsClassifier	0.69	0.63	0.65	0.61
SVC	0.69	0.60	0.56	0.63
NuSVC	0.70	0.63	0.65	0.62
LogisticRegression	0.71	0.61	0.57	0.66
DecisionTreeClassifier	0.74	0.67	0.67	0.67
RandomForestClassifier	0.76	0.71	0.71	0.71
CatBoostClassifier	0.79	0.74	0.74	0.73
XGBClassifier	0.81	0.76	0.75	0.77

RESULTS

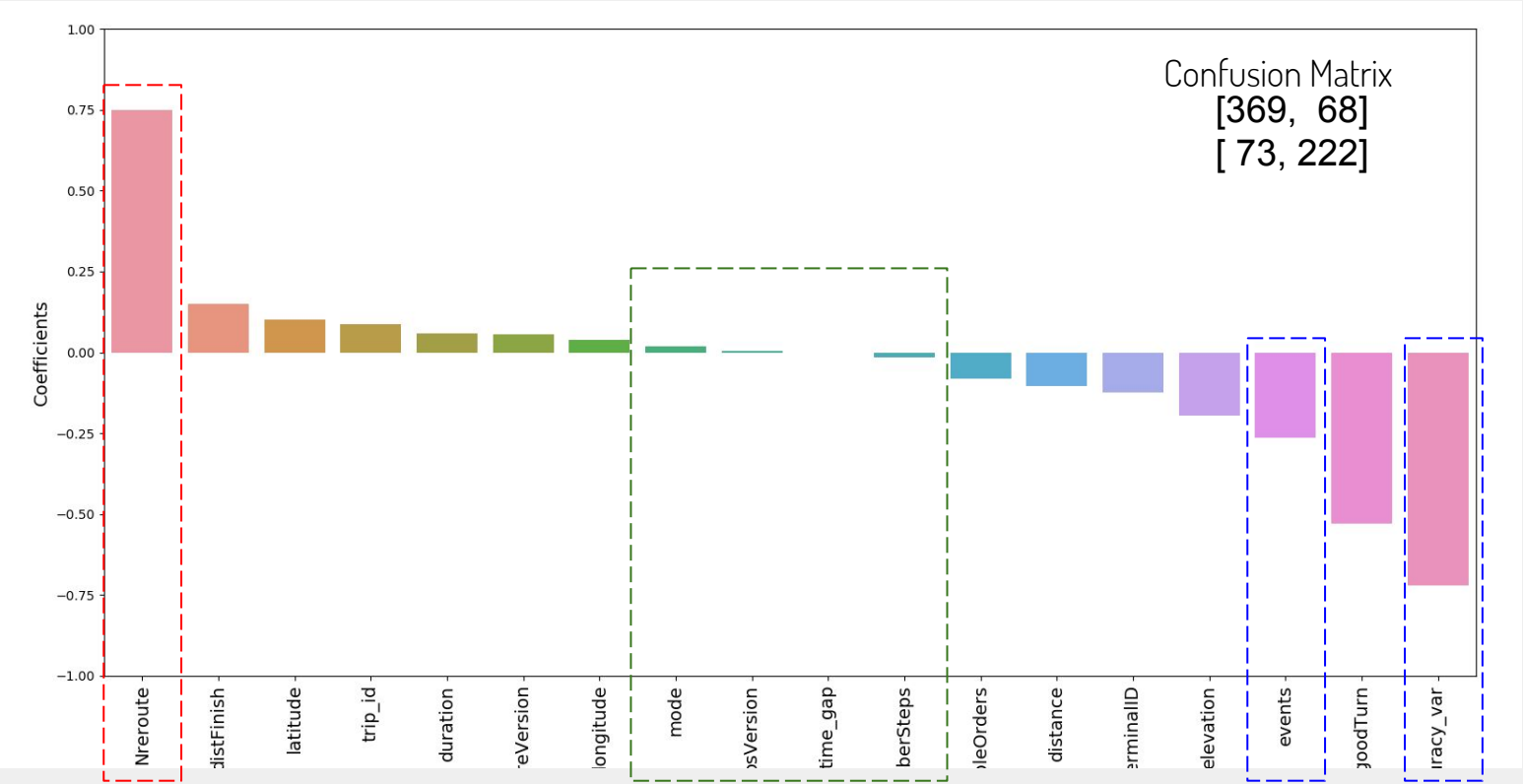
Use N of reroute to the moment of turn
Same for good turn

Logit Regression Results

Dep. Variable:	badturn	No. Observations:	2216
Model:	Logit	Df Residuals:	2209
Method:	MLE	Df Model:	6
Date:	Fri, 06 Mar 2020	Pseudo R-squ.:	0.1346
Time:	12:29:25	Log-Likelihood:	-1292.9
converged:	True	LL-Null:	-1494.0
Covariance Type:	nonrobust	LLR p-value:	9.107e-84

	coef	std err	z	P>|z|	[0.025	0.975]
events	-0.5534	0.051	-10.947	0.000	-0.653	-0.454
elevation	-0.2246	0.050	-4.516	0.000	-0.322	-0.127
accuracy_var	-0.2657	0.077	-3.440	0.001	-0.417	-0.114
terminalID	-0.2061	0.052	-3.989	0.000	-0.307	-0.105
distFinish	0.2711	0.054	4.982	0.000	0.164	0.378
Nreroute	0.7308	0.059	12.345	0.000	0.615	0.847
NgoodTurn	-0.4678	0.058	-8.033	0.000	-0.582	-0.354

RESULTS


Confusion Matrix
[369, 68]
[73, 222]
Coefficients
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Nreroute
latitude
trip_id
duration
longitude
mode
time_gap
distance
elevation
events

RESULTS SUMMARY

- **What is the "Value" of Data (if there is enough)**

  More better data - More users sending data to improve model but better selecting features to store

- **What is the Reroute event rate?**

  With the current data rate is around 40% but there are not enough data just for KPI per trip

- **What are the reroute events situation ?**

  As suggested by the models GPS accuracy variation and Type of Turns (need more details)

- **What are the relevant Navigation statistics?**

  Number of rerouting, Distance to Finish, Trip Duration, trip Distance

- **How much improvement in operation can be achieved ?**

  Just by processing the data prior to storage reduce 18% of data files

FUTURE

More UserData Is needed

Automating and Optimizing the KPIs script

Storage

- Map tool parameters (firmware change)
- Cloud computing
- Less data but more relevant

Model

- Improve cleaning (speed, direction, turn direction)
- Hyperparameter tuning

KPIs and additional features

Actual route vs Planned Route and Add crossroad types

Mapbox query against Google Maps