Ninth Montreal Industrial Problem Solving Workshop: Spilling problem

Ismael Assani, Poclaire Kenmogne, Jiliang Li, Gabriel Lemeyre, Thi Thanh Hue Nguyen, Frédérique Robin, Pierre-Loïk Rothé

August 26, 2019

Under the supervision of
François Bellavance (HEC) & Olivier G. Leblanc (Air Canada)

AIR CANADA

Outline

Context

The objective is to predict the spilling of a flight:
$\hookrightarrow$ Spilling flight definition: open to interpretation.

One proposition: the spill flight event is defined by the event:

$$\text{“occupation rate 3 days before departure} \geq 0.95\text{”.}$$

The occupation rate is defined by

$$\frac{\text{the number of bookings}}{\text{actual airplane capacity}}.$$

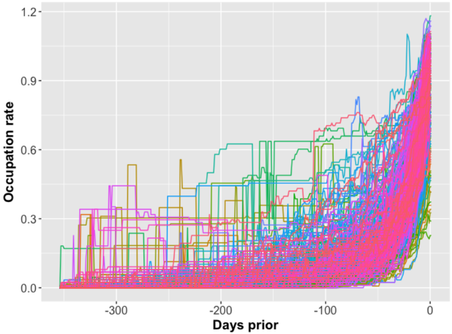
Context

One proposition: the spill flight event is defined by the event:

$$\text{“occupation rate 3 days before departure} \geq 0.95\text{”}.$$

The occupation rate is defined by

$$\frac{\text{the number of bookings}}{\text{actual airplane capacity}}.$$



Dataset Building and Features - 1

10 Origins and Destinations: AAA, BBB, CCC, ..., JJJ, KKK, LLL

AAA-BBB
BBB-AAA, BBB-DDD, BBB-EEE, BBB-III, BBB-JJJ BBB-LLL
CCC-FFF
DDD-BBB
EEE-BBB
FFF-CCC, FFF-GGG, FFF-HHH
GGG-FFF, GGG-KKK
HHH-FFF III-BBB
JJJ-BBB KKK-GGG LLL-BBB

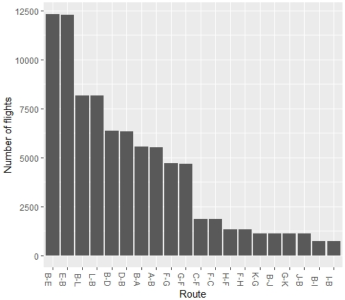


Figure: Flights from 20 routes studied between 10 Origins and Destinations

Simplification: aggregating data by a unique flight index (TOD).

$\hookrightarrow$ longitudinal data (time series) per each flight over two years.

Dataset Building and Features - 2

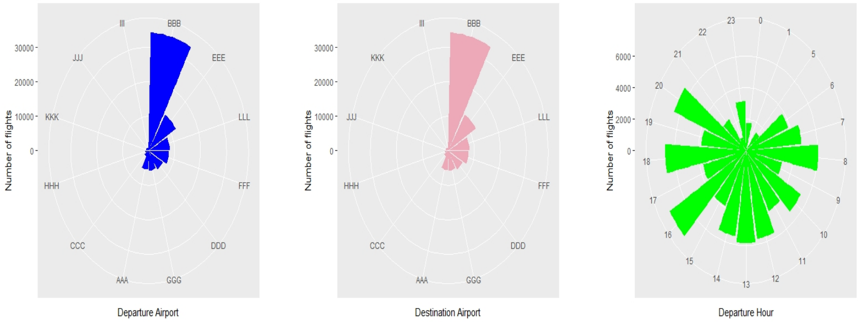
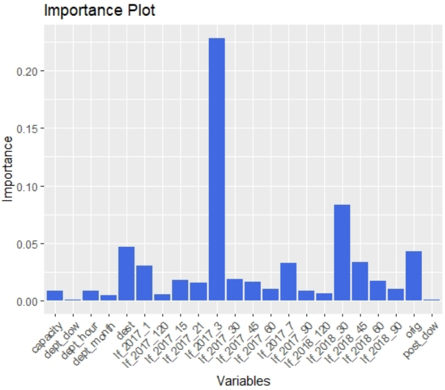


Figure: Distribution of flights functions of Departure airport (Left), Destination airport (Middle) and Departure hour (Right)

Machine Learning - Random forest - 1

Machine Learning - Random forest - 2



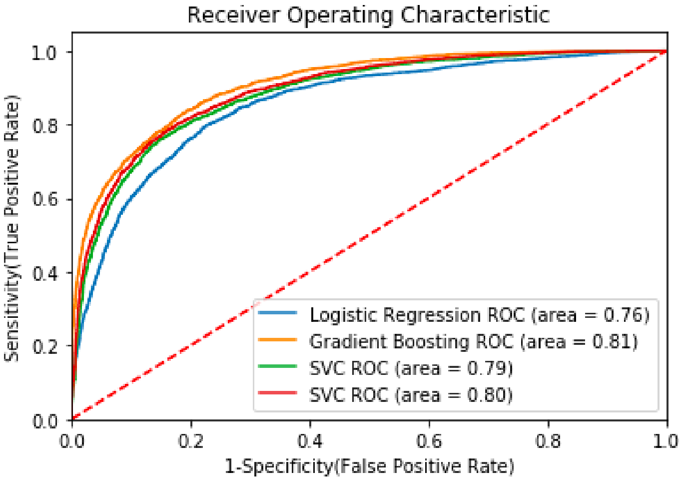
- Data were stratified by cabin class
- 70-30 split between training and testing data
- 5-fold cross-validation using *caret* package
- Average of 93% of accuracy achieved for spill-detection

Machine Learning - Lasso, SVM, Gradient Boosting, logistic regression

	predictors	coefficients	sort
7	VRTUL_CAP_CNT	1.992305	1.992305
6	ADJ_CAP_CNT	-1.103309	1.103309
5	PHY_CAP_CNT	-1.037746	1.037746
15	LF_2018_45	0.198183	0.198183
23	LF_2017_1	0.172539	0.172539
35	LF_2017_45	-0.102602	0.102602
22	cap_2017_1	0.101714	0.101714
3	CARRIER_CDE	0.036218	0.036218
14	cap_2018_45	0.032820	0.032820
9	OVR_BOOK_FARE_CAD	-0.024283	0.024283
4	FLIGHT_NUM	0.024120	0.024120
0	Online.Path	-0.022669	0.022669
8	OVR_BOOK_SEAT_CNT	0.022425	0.022425
29	LF_2017_15	-0.021366	0.021366
25	LF_2017_3	0.015092	0.015092
1	TOD	-0.013028	0.013028
41	LF_2017_120	0.008564	0.008564
17	LF_2018_60	-0.008296	0.008296
19	LF_2018_90	-0.007817	0.007817

- Data were stratified by flight
- We use Lasso to select features.
- Average of 80% of accuracy achieved for spill-detection for SVM, LG, Gradient Boosting

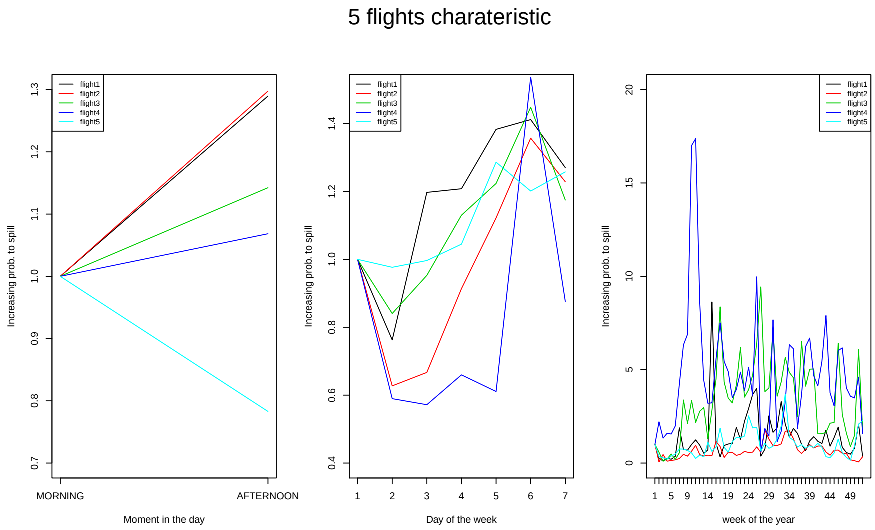
Machine Learning - ROC , AUC



Survival model approach - 1

- Approach: Train a survival model to obtain a survival function associated with each unique Origin-Destination pair.
- The selected model is the Cox.
- This model allows us to predict the probability of survival according to certain flight characteristics.
- The characteristics retained are: the moment of the day, the day of the week and the week of the year.

Survival model approach - 2



Survival model approach - 3

Application to one flight to predict the probability of spill 3 days before departure knowing that we are 30 days from departure gives : prediction score = 67.01%; MSE = 53.17%.

- Low prediction capacity: But normal since the model does not take into account any other information.
- Can be use as feature engineering to improve another model.
- Possible improvement : add more relevant variables that may explain spill (eg: price range 30 days before departure).

Kalman Filtering - 1

Approach: Compute a forecast of plane occupation and conclude if it spills or not

Historical data and measurements $\longrightarrow$ Occupation rate forecast $\longrightarrow$ Spilling Forecast

Principle:

- Infer dynamic for the current booking:
  Use historical data to fit a polynomial regression
- Modify dynamic to fit current measurements (data-driven approach):
  Use Kalman filtering to enrich the dynamic with current observations

Kalman Filtering - 2

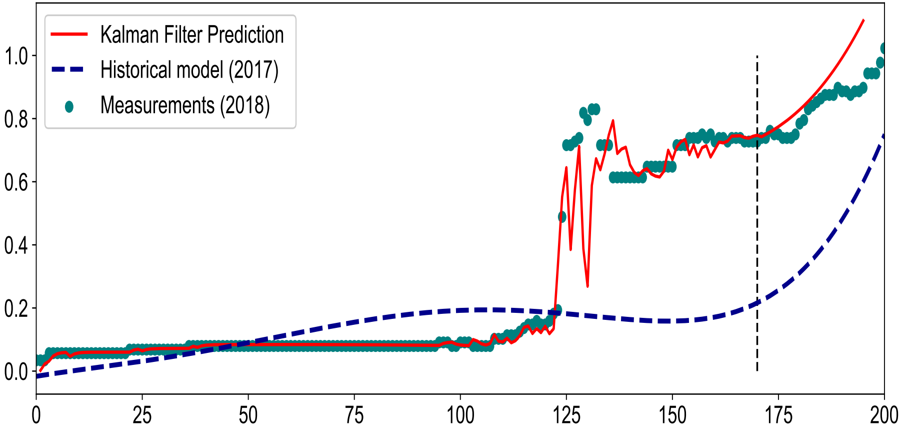


Figure: One flight occupation prediction using Kalman Filters and historical model (Polynomial degree: 5)

Kalman Filtering - 3

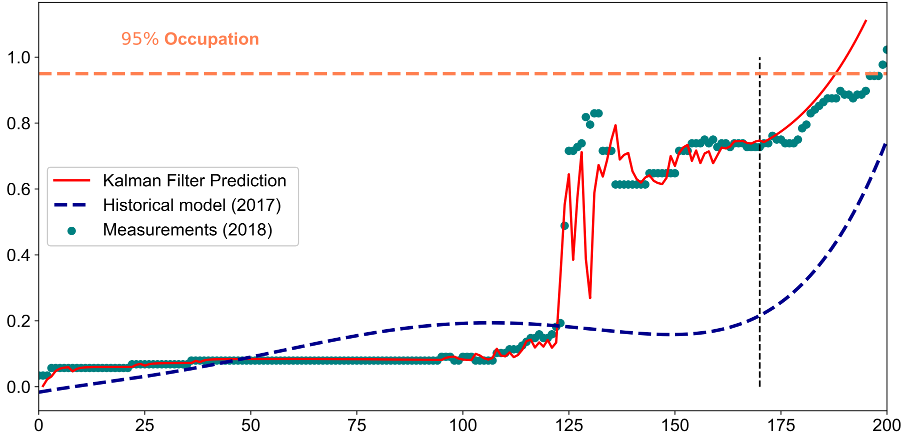


Figure: One flight occupation prediction using Kalman Filters and historical model (Polynomial degree: 5)

Kalman Filtering - 4

	Actual	Predicted
Spill occurrence rate	36%	40%

Figure: Results for a dataset of 11,307 flights

- Prediction score: 73%
- False negative: 12%

Perspectives:

- Improving the historical dynamic model
- Machine learning initial guess for new flight (without historical data)
- The Kalman filtering approach allows day to day update of the occupation forecasting with minimal computational load

Acknowledgments

Thank for your attention !

Special thanks to Olivier, François, Caroline and Fabian . . .

and Odile for the organization