Prédiction de l'occurrence d'une baisse de prix pour le conseil à l'achat d'un billet en ligne

Till WOHLFARTH[1,2], Stephan CLEMENÇON[2], François ROUEFF[2], Xavier CASELLATO[1]

[1]liligo.com
4, Allée verte, 75011 Paris, France

[2]Telecom Paristech- LTCI (Institut Telecom - CNRS)
46 rue Barrault, 75634 Paris cedex 13, France

till@liligo.com, stephan.clemencon@telecom-paristech.fr
francois.roueff@telecom-paristech.fr xavier.casellato@liligo.com

Résumé – Nous nous intéressons au problème de la prédiction de l'occurrence d'une baisse de prix pour fournir un conseil à l'achat immédiat ou reporté d'un voyage sur un site web de comparaison des prix. La méthodologie proposée repose sur l'apprentissage statistique d'un modèle d'évolution du prix à partir de l'information conjointe d'attributs du voyage considéré et d'observations passées du prix et de la "popularité" celui-ci. L'originalité principale consiste à représenter l'évolution des prix par le processus ponctuel inhomogène des sauts de celui-ci. Cette représentation servira en particulier à classer les vols (issus de la base de données de liligo.com) en comportements similaires et à construire pour chacun des comportements identifiés un modèle commun. Nous mettons alors en oeuvre une méthode d'apprentissage d'un modèle d'évolution des prix. Ce modèle permet de fournir un prédicteur de l'occurrence d'une baisse du prix sur une période future donnée et donc de prodiguer un conseil d'achat ou d'attente au client. L'approche par modèle est comparée à une approche directe.

Abstract – The goal of this paper is to consider the design of decision-making tools in the context of varying travel prices from the customer's perspective. Based on vast streams of heterogeneous historical data collected through the internet, we describe here two approaches to forecasting travel price changes at a given horizon, taking as input variables a list of descriptive characteristics of the flight, together with possible features of the past evolution of the related price series. Though heterogeneous in many respects (*e.g.* sampling, scale), the collection of historical prices series is here represented in a unified manner, by *marked point processes* (MPP). State-of-the-art supervised learning algorithms, possibly combined with a preliminary clustering stage, grouping flights whose related price series exhibit similar behavior, can be next used in order to help the customer to decide when to purchase her/his ticket.

1 Introduction

Motivations et objectifs Les compagnies aériennes au premier chef, suivies par l'ensemble des professionnels du tourisme (compagnies ferroviaires, hôteliers, etc.) ont généralisé les politiques de "yield management" afin d'optimiser le prix d'une prestation en fonction de leur niveau d'inventaire et de la date de réservation [6]. Il en résulte une opacité totale dans le processus de formation des prix, qui, pour un même billet et aux mêmes dates, peuvent varier très fortement (i) d'un fournisseur à un autre, et (ii) d'un moment à l'autre. Le consommateur est maintenu dans l'ignorance et l'incertitude, parfois encouragé à réserver longtemps à l'avance, parfois exhorté à réserver sur un coup de tête à la dernière minute pour bénéficier d'offres présentées comme dégriffées.

Liligo.com est un moteur de recherche du voyageur capable de chercher un billet d'avion parmi plus de 250 sites d'agences de voyages et compagnies aériennes. Afin d'aider l'utilisateur dans son acte d'achat, nous voudrions pouvoir afficher pour chaque vol retourné par la recherche, une aide à la décision d'achat basé sur une estimation de la tendance dans l'évolution du prix.

Base de données Liligo.com possède une base de données d'historiques de recherche des utilisateurs sur laquelle nous avons constitué notre base d'apprentissage et de test. Dans cette base de données, un *trajet unique*, correspondant à un résultat de recherche, est défini par 6 attributs : l'aéroport de départ, l'aéroport d'arrivée, le date de départ, la date de retour, le code transporteur de vol allé et le celui du vol retour. Cela implique que pour un même trajet, on peut avoir autant de séries temporelles que de sites proposant ce vol (site de la compagnie aérienne mais aussi sites d'agence de voyages).

Pour notre étude, nous nous sommes concentrés sur 6 trajets provenant de 9 sites différents (voir Tableau ci-contre). Nous étudions des trajets aller-retour (représentant 80% des demandes) pour des séjours de 3, 7 et 14 jours afin de couvrir la majorité des recherches. Le choix des destinations s'est faite afin d'optimiser l'échantillonnage des courbes : ce sont des trajets suffisamment demandés pour avoir un prix environ toutes

From	To	Provider	Durée
Paris	Budapest	Malev	3,7
Paris	Budapest	EasyJet	3,7
Paris	Toulouse	EasyJet	3,7
Paris	Toulouse	iDTGV (Train)	3,7
Paris	Marrakesh	Transavia	3,7
Paris	Bangkok	Quatar	7,14
Amsterdam	Barcelona	Austrian	3,7

TABLE 1 – Trajets considérés et durées de séjour utilisés dans la base de données.

les 6 heures.

Pour chaque trajet unique, nous collectons un ensemble d'attributs $V_i(1), \ldots, V_i(p)$ qu'on peut répartir en trois catégories : les caractéristiques du vol (lieu de départ, lieu d'arrivée, numéros des vols, type d'avion, durée du séjour), les attributs temporels (horaires de départ et de retour, jour de la semaine, jour de l'année, saison, période de la journée) et les autres (nombre de recherches pour ce billet, site d'origine du prix).

1.1 Modélisation des courbes de prix

Pour chaque élément i de la base de données, on dispose d'une courbe $p_i(t)$ représentant le prix d'achat du voyage à l'instant t. Ces points sont interpolés afin d'obtenir une courbe constante par morceaux. Les sauts sont produits par l'application d'une règle de "yield management" qui reste cachée aux consommateurs. On introduit alors les instants de sauts $T_k^{(i)}$ numérotés dans l'ordre croissant de telle sorte que $T_0^{(i)}$ est la date de départ (les indices k sont donc négatifs). On suppose par convention que le prix est continu à droite, d'où la courbe de prix interpolée définie pour tout $t < T_0^{(i)}$ par :

$$p_i(t) = \sum_{k \leq 0} p_i(T_{k-1}^{(i)}) \mathbb{1}_{[T_{k-1}^{(i)}, T_k^{(i)})}(t) \ .$$

On définit alors la suite des rendements, ou *taille relative des sauts* :

$$s_k^{(i)} = \{p_i(T_k^{(i)}) - p_i(T_k^{(i)}-)\}/p_i(T_k^{(i)}-), \quad k \leq 1 \ ,$$

où $p_i(t-)$ désigne le prix juste avant l'instant t. Il est clair que la courbe des prix peut être entièrement reconstruite à partir du prix initial et de la suite des points $(T_k^{(i)}, s_k^{(i)})$. Nous modélisons donc cette suite de points par un processus ponctuel marqué inhomogène, voir [3], dont l'intensité $J(t,s)$ peut être estimé estimé sous la forme d'une image pixélisée qui prend les valeurs

$$\widehat{J}_i(s,t) = \frac{1}{b_1 b_2} \sum_{k \leq -1} 1_R(T_k^{(i)}, s_k), \quad (s,t) \in R \ , \qquad (1)$$

pour un pixel rectangulaire R de taille $b_1 \times b_2$. Le plan temps /rendement est partitionné par une grille régulière de tels pixels. Et à chaque pixel est donc associée une intensité égale au nombre moyen de sauts par unité de surface dans le plan temps /rendements à l'intérieur du pixel. Notons que pour chaque vol nous ajoutons à la liste des attributs $V_i(1), \ldots, V_i(p)$, la somme des valeurs de chaque pixel précédant la prédiction, ainsi que tous les pixels en tant qu'attributs. Nous aurons ainsi une information sur la volatilité des premiers prix ainsi qu'une information plus fine sur le comportement initial du vol. Les trois étapes de cette représentation sont illustrées en Fig. 1, 2 et 3, pour un vol Paris-Marrakech sur une compagnie française low cost pour un départ le 24 octobre 2010.

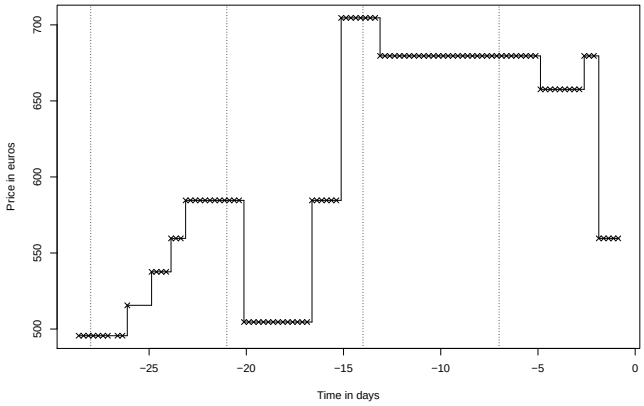

FIGURE 1 – **Série temporelle : points observés ('*'), courbe interpolée (trait plein), les lignes verticales indiquent in changement de semaine.**

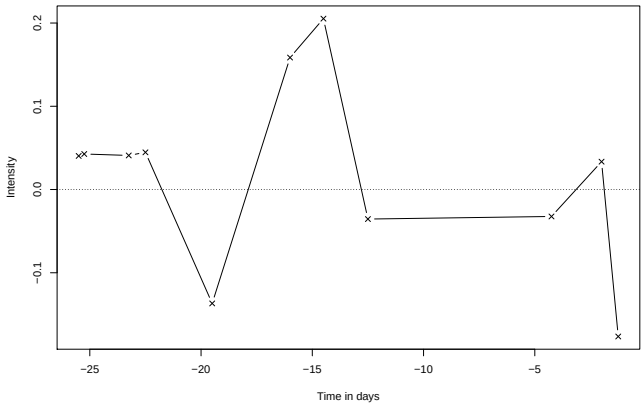

FIGURE 2 – **Série de sauts relatifs : processus marqué des points des sauts de la courbe interpolée des prix.**

2 Méthodologie

Supposons qu'un utilisateur soit prêt à acheter un voyage correspondant au trajet i t_1 jours avant la date de départ, soit à la date $T_0^i - t_1$. Nous proposons une méthode pour prédire l'évolution future du prix $p_i(t)$ pour $T_0^i - t_1 < t < T_0^i$, à

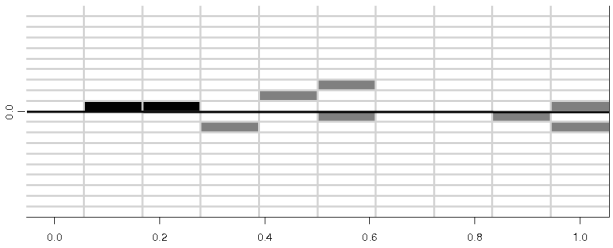

FIGURE 3 – **Estimateur de l'intensité du processus de point sous la forme d'une image pixélisée pour un résolution donnée par $b_1 = 3\ jours$ et $b_2 = 0.1$.**

partir de ses attributs (compagnie,horaires,...) et des premiers prix collectés. Pour cela la base d'apprentissage est utilisée 1) pour construire un nombre fini C de modèles d'évolution du prix par un algorithme d'apprentissage non–supervisé 2) pour apprendre le meilleur modèle à partir des attributs du trajet par un algorithme de classification supervisé. Le modèle obtenu est alors utilisé pour prédire l'évolution du prix.

Construction de modèle par apprentissage non–supervisé Dans un premier temps, nous créons des groupes ayant des comportements similaires en appliquant l'algorithme de classification des KMeans [4] en se basant sur les images pixélisées d'intensité $\widehat{J}_i$ pour i parcourant la base d'apprentissage. On obtient ainsi C classes d'indices $I_1, \ldots, I_C$ permettant de regrouper chaque item i de la base d'apprentissage par comportements similaires d'évolution du prix. Le nombre optimal de classes est choisi en appliquant la méthode Gap [7] sur notre base d'apprentissage. Cette étape est primordiale dans la construction de notre modèle, un mauvais choix de C pouvant dégrader rapidement l'étape de clustering et donc la prédiction. Pour chaque classe $j = 1, \ldots, C$, un processus ponctuel aléatoire est défini pour les points (T_k, s_k) et fournit un modèle commun aux éléments de la classe.

Classification par les attributs Nous construisons un classifieur qui prend en entrée un trajet i et son vecteur d'attributs $V_i(1), \ldots, V_i(p)$, ainsi que les premiers prix observés pour lui assigner un numéro de groupe $j \in \{1, \ldots, C\}$. Nous utiliserons deux algorithmes d'apprentissage supervisés, à savoir, l'arbre de classification (CART [2]) et le random forest [1].

L'arbre de classification nous permet d'observer les règles de classification créées pour en donner directement une interprétation et repérer les attributs importants. Les random forest construisant une multitude d'arbres nous permettra aussi de classer chaque attribut par ordre d'importance dans la classification.

Prédiction d'une baisse des prix Posons par commodité la date de départ à l'origine, $T_0^i = 0$. Nous définissons la variable $\varphi_i \in \{0, 1\}$ correspondant respectivement aux conseils "achat" et "attendre". Partant du principe que le client est capable de vérifier les prix une fois par jour jusqu'à $-t_2 \in (-t_1, 0]$, nous définissons $\varphi_i = 1$ si et seulement si le prix $p_i(t)$ reste inférieur à $p_i(-t_1)$ pendant plus de 24 heures entre $-t_1$ et $-t_2$. Une fois la classe j obtenue en appliquant le classifieur aux attributs du trajet i considéré, le modèle de la classe j est utilisé pour calculer la probabilité $\mathbb{P}(\varphi_i = 1)$. Cette valeur est utilisée pour proposer un conseil d'achat ou d'attente pour un niveau de confiance donné.

3 Résultats

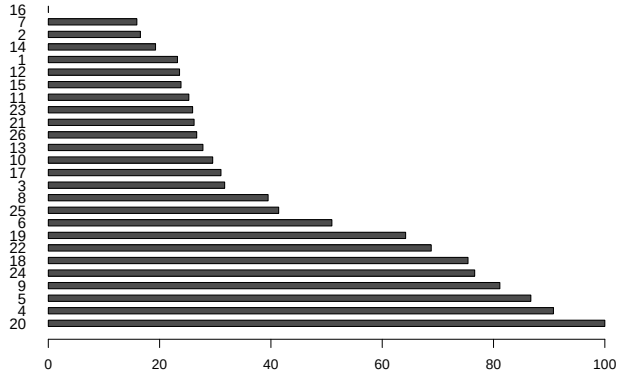

FIGURE 4 – **Tableau d'importance relative des attributs**

Importance relative des attributs Pour des méthodes de classification telles que CART ou random forest (voir [1]), il est possible de classer les attributs selon leur importance relative dans la règle prédictive (Figure 4). Les attributs temporels tels que le jour de l'année ou le jour du mois, permettent une meilleure segmentation des vols en groupes, validant les phénomènes de saisonnalité. La somme des demandes passées apparaît, dans le tableau relatif à la règle produite par l'algorithme random forest, en troisième position, ce qui confirme l'importance du principe de l'offre et de la demande l'évolution des prix. Enfin, le nombre de sauts précédant la prédiction ainsi que les pixels de faible rendement font partie des attributs ayant le plus d'impact.

Sur la Figure 5, on observe la corrélation partielle de la variable "Jour de l'année", décrivant la dépendance marginale de cette dernière dans la prédiction (voir [5]). L'effet des périodes de vacances et de fêtes y est clairement visible.

Performances Nous avons évalué la méthode proposée à partir de la base de test. Elle est comparée à une approche directe pour laquelle la variable φ_i est directement prédite à partir des attributs du vol par un apprentissage supervisé appliqué à la

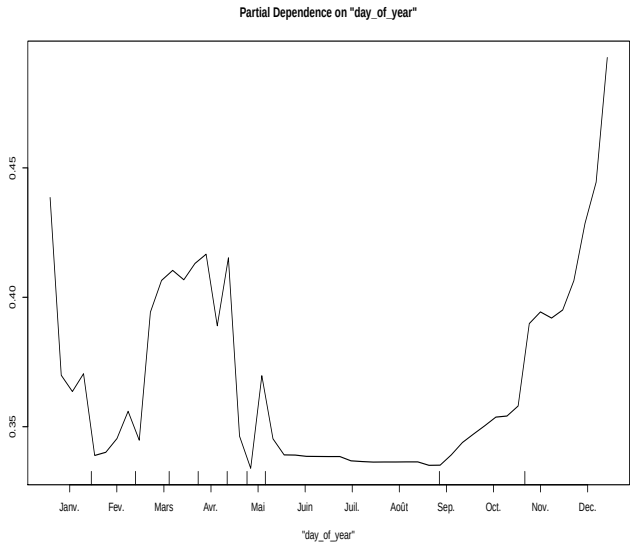

FIGURE 5 – **Corrélation partielle de la variable “Jour de l’année”**

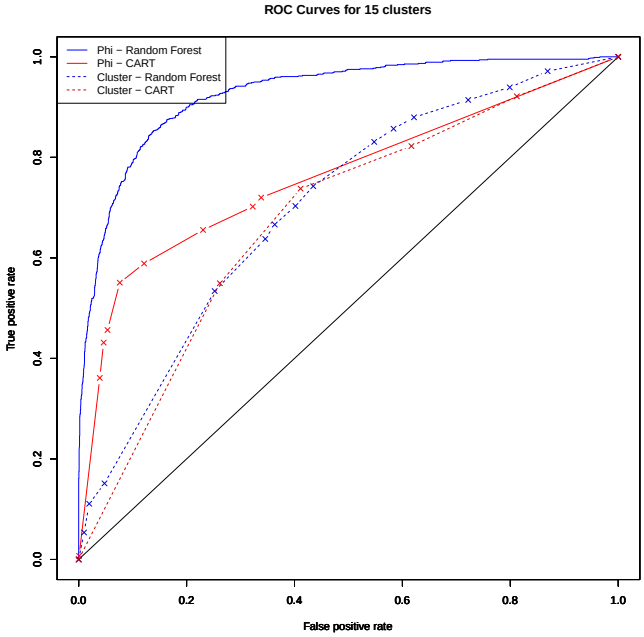

FIGURE 6 – **Courbe ROC : performance de l’approche directe (Phi) et de l’approche par les clusters (Cluster) par CART (rouge) et par Random Forest (bleu)**

base d’apprentissage. Cette méthode est par nature plus efficace puisqu’elle a pour seul objectif d’optimiser la prédiction de φ et non d’apprendre un modèle complet de l’évolution du prix du trajet.

Notre mesure de performance est le “receiver operating characteristic” [8] (courbe ROC, Figure 6). Cela nous permet de comparer l’approche directe à celle par modèle ainsi que les deux algorithmes de classification utilisés. L’approche directe donne bien de meilleurs résultats. Néanmoins un modèle permet de s’adapter aisément aux contraintes du client pour l’achat du billet. En particulier la définition de la variable φ dépend des contraintes sur les dates d’achat et sur la fréquence de visite de l’utilisateur du site pour surveiller une baisse du prix. Chaque définition de φ différente nécessite une nouvelle phase d’apprentissage alors que si un modèle est d’ores et déjà disponible, la valeur $\mathbb{P}(\varphi_i = 1)$ est calculable sans renouveler l’étape d’apprentissage.

4 Conclusion

En conclusion, l’introduction d’une nouvelle représentation de séries temporelles permet ici une meilleure agrégation et l’élaboration d’un modèle de probabilité. Nous avons ainsi réussi (i) à prédire précisement de manière directe l’évolution du prix d’un trajet et (ii) à modéliser des comportements généraux nous permettant par la suite de prédire les évolutions possibles. Nous comptons maintenant améliorer l’étape de clustering en remplassant les k-means par un modèle de mélange utilisant un algorithme Expectation-Maximization. Cela permettra une estimation et une interprétation plus rigoureuse de l’approche (ii). La mise en place de ce service sur le site liligo.com nécessitera de plus de faire des mises à jour en temps réél de nos résultats en fonction de l’augmentation de la base de donnée disponible et une meilleure prise en compte des premiers points et de l’évolution de la demande.

Références

[1] L. Breiman. Random Forests. *Machine Learning*, 45(1) :5–32, October 2001.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[3] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2003. Elementary theory and methods.

[4] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data : An Introduction to Cluster Analysis*. Wiley, 2005.

[5] M. Kendall and A. Stuart. *The advanced theory of statistics*, volume 2 : Inference and Relationship. Charles Griffin, London, 4th edition, 1979.

[6] B. Smith, J. Leimkuhler, R. Darrow, and Samuels. Yield management at American Airlines. *Interfaces*, 22(1) :8–31, 1992.

[7] R. Tibshirani, G. Walther, and T. Hastie. Estimating the Number of Clusters in a Dataset via the Gap Statistic, 2000.

[8] H. van Trees. *Detection, Estimation, and Modulation Theory*, volume 1. Wiley, New York, 1968.