Things Change: Comparing Results Using Historical Data and User Testing for Evaluating a Recommendation Task

Soon-Gyo Jung
Qatar Computing Research Institute
HBKU, Doha, Qatar
sjung@hbku.edu.qa

Joni Salminen
University of Turku, Turku, Finland; Qatar Computing Research Institute
HBKU, Doha, Qatar
jsalminen @hbku.edu.qa

Shammur A. Chowdhury
Qatar Computing Research Institute
HBKU, Doha, Qatar
shchowdhury@hbku.edu.qa

Dianne Ramirez Robillos
School of Statistics
University of the Philippines
Quezon City, Diliman, Philippines

Bernard J. Jansen
Qatar Computing Research Institute
HBKU, Doha, Qatar
jjansen@acm.org

CHI'20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA

ACM ISBN 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.3382945

Abstract

We address a recommendation task for next likely flight destination to customers of a major international airline company. We compare performance using historical flight data and an actual user evaluation. Using two years of historical flight data consisting of tens of millions of flights, an ensemble and a collaborative filtering approach obtained an accuracy of 47% and 20% using a test set of 100,000 customers, respectively, highlighting the challenge of the domain. We then evaluated our recommendations on 10,000 actual customers, with a 45-45-10 split among ensemble, collaborative filtering, and control group. The overall predictive power employed with real users was 23%, with the ensemble method having a predictive power of 19% and 30% for collaborative filtering. Results indicate that, in complex and shifting domains such as this one, one cannot rely solely on historical data for evaluating the impact of user recommendations. We discuss implications for recommendation systems and future research in this and related domains.

Author Keywords

Prediction; Recommendations; Algorithmic trade-off; User Study

Ensemble approach - an approach of using of multiple machine learning algorithms with the aim of obtaining better predictive performance than one could obtain from using any single machine learning.

Collaborative filtering - an approach of employing user attributes for identifying (i.e., filtering) items that a user might like based the reactions of users with similar attributes.

Sidebar 1: Definitions of two algorithmic approaches used to analyze historical data.

CSS Concepts

• Human-centered computing → Human computer interaction (HCI); HCI design and evaluation methods;

Introduction

There are situations where one wants to predict what user action (i.e., state) will come next in a sequence. Examples include the next button a user pushes on a system, a query that a user submits to a search engine, or a result that a searcher next clicks on a results listing page. In these situations, predictive approaches typically employ historical data to forecast where a user will go next in a sequence.

A related situation is where one is not strictly predicting the next state but recommending actions that the user might not have done before but would prefer to take. Examples of this could be recommending a feature of a program that a user may not have tried before, a book to read, or a movie to watch. These recommendation approaches typically focus on novelty based on previous behavior or recommending actions at an appropriate place in a given sequence.

However, there is a third context that is a blend of these two. These are situations where one wants to predict someone will return to a prior state but also recommend new states that are similar to but not the exact states previously visited. Flight booking is one of these contexts. Flight bookings is a particularly challenging domain, as, in addition to a blended prediction–recommendation context, there are many confounding factors such as business versus leisure travel, flight time, vacation, costs, etc. that may impact the customer's final decision concerning the suggestion.

It is this challenging flight booking domain that is the focus of our research, with a specific interest on the effectiveness of predicting user behaviors within this blended context. Our research goal is to *investigate the effectiveness of algorithmic approaches for suggesting next flight destinations for customers.*

To do so, we implement two algorithmic approaches, an ensemble approach and a collaborative filtering approach that are both well-known algorithmic methods in the recommender domain [e.g., 13], to evaluate their effectiveness for recommending the next likely flight destinations for customers. We first train and afterward test our models using historical customer flight data. Then, we conduct a user evaluation [7] using 10,000 customers of a major international airline to recommend a next likely destination booking. We implement the designed ensemble and collaborative filtering approaches in a 45-45-10 (control group) for the user study.

Results show that the accuracy of the recommendation algorithms was substantially different between using just historical data and employing the approaches with actual users. The implication is that these hybrid recommendation contexts require validation with actual users, rather than relying on solely historical data, in order to get a true measure of the predictive accuracy and algorithmic effectiveness.

Prior Work

Recommendation approaches provide suggestions that users might be interested in by analyzing, typically, past behavior of users in order to build a profile of interests or behaviors, and then leverage this profile in order to recommend potential future states (i.e.,

features, destinations, movies, etc.) with the standard approach modeling a series of actions as states [16].

Such an approach can be highly accurate if the task is one of algorithmic prediction (i.e., predicting the next state), including re-visitation and reuse [4]. However, a known limitation of this approach when actually deployed with users is that the suggestions may be states that are nearly identical to what the user knows already. Additionally, there are contexts where one may not want to predict as much as recommended in a serendipitous manner new states that the user probably will like but may not be aware exists.

This content is especially applicable to the domain of flight bookings, which is both a prediction (i.e., where will the passenger next book) and recommendation (i.e., where would the passenger like to book) problem. In the prediction aspect of the problem, re-visitation to a prior state is acceptable. In the recommendation aspect, the suggested state is a novel, one.

There has been prior work concerning the issue serendipity in recommender systems, including the discovery of similar users based on their temporal histories [3] and various tactics for personalized user profiling [14]. There have been various algorithmic approaches including classes of collaborative filtering and ensemble approaches [16]. Findings show that combining multiple approaches is not always more effective than using single a single approach in terms of average prediction accuracy and that in given contexts, such as temporal, different approaches perform better or worse [16]. In the competitive airline industry, passenger prediction and recommendation are active avenues of pursuit aimed at in-depth insights in customer behaviors [6], generating revenue [12], and enhancing the customer experience [2]. Approaches taken have focused on both aggregated data and individual user data [8].

However, most of the prior work in the prediction-recommendation area have been developed and evaluated on historical datasets using testing and training portions and not deployed with real people who then react to the suggestions. This limitation is especially problematic in the airline domain, where selecting a destination can be impacted by a variety of confounding factors. For example, a customer may be receptive to a recommended destination but have a business trip that is out of the customer's control. There are also departure times, flight durations, costs, and many other constraining factors on recommendations. Some of these are also, of course, present and have been noted in other contexts [1, 5, 11]. These confounding factors raise concerns of using historical data to evaluate algorithmic accuracy.

Therefore, there are several unanswered questions. *How does the algorithmic accuracy of historical data compare to that with real users? How do recommendation algorithms perform with real people for flight bookings?* These are some of the questions that motivate our research.

Research Objectives

Our research objectives are:

a. Develop a next likely destination recommendation approach for airline customer bookings,
b. Measure the accuracy of the model using historical data,

Component	Definition
Customer Retention Program	Activities that are taken to reduce the number of customer defections. The goal of customer retention programs is retaining as many customers as possible
High Customer Retention	Customers return to, continue to buy or not defect to another business or to non-use entirely
Customer Retention Rate	Percentage of customers the company has retained in a given period

Table 1: Key Components of and definitions of customer retention programs.

c. Evaluate the accuracy of the model via a marketing campaign in a commercial setting using real users
d. Compare the evaluation results using historical data with using booking from actual users.

As part of a customer retention initiative at a major international airline, the organization desires to analyze historical flight data from customers to suggest a next likely destination (NLT) from which the organization will then send an incentive offer to these customers for this destination in order to lock in the booking and reduce the chance of the customer going to another airline for the booking. See Table 1 for key constructs and definitions of customer retention.

Methodology

For the algorithmic approaches, we implement two methods, an ensemble and collaborative filtering algorithms. The flight data used historical data for each customer (bookings, dates, destinations, ticket class, gender, age, etc.). This historical data was used as input for the ensemble and collaborative filtering methods to develop a next likely destination suggestion for these airline customers. The overall methodological concept is shown in Figure 1.

Tested Algorithmic Approaches

For the first tested approach, the ensemble method, merges several classifiers together with the aim of achieving better performance than any single classifier. The underlying assumption is that different classifiers have different strengths and weaknesses. Ensemble methods are known to decrease variance, decrease bias, or improve predictions [10].

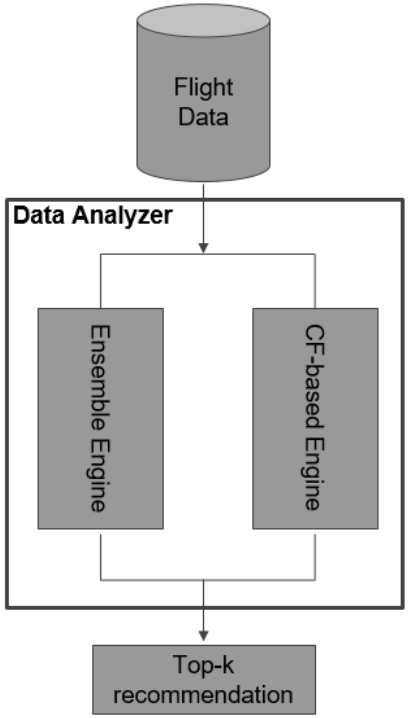

Figure 1: Methodological framework for NLT A/B testing.

The best combination of classifiers was Bagging, Multilayer perceptron, and KNN, which is the combination that we employed in this research. We tested several classifiers and compared them based on accuracy. Classifiers considered were Logistic Regression, Naive Bayes, Ada Boosting, Bagging, Random Forest, Gradient Boosting, Support vector machine (SVM), Multilayer perceptron, K-nearest neighbors (KNN), and Quadratic discriminant analysis. These algorithms were chosen because they represent state-of-the-art in many recommendation approaches

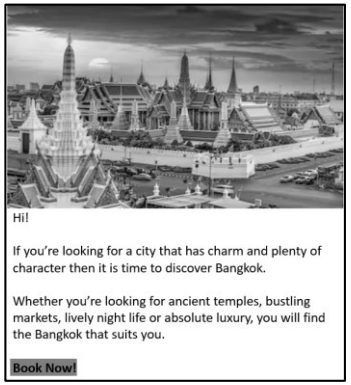

Figure 2: Bangkok marketing message sent to the test customers. Altered to remove branding.

Figure 3: London marketing messages sent to the test customers. Altered to remove branding.

The second approach was collaborative filtering that finds similar users to a given user and then recommends recent destinations of these similar users to the given user. The underlying assumption is that similar individuals will be interested in similar destinations. So, if person A has flown to the same destination as person B, person A is likely to also like B's other destinations than destinations of a random person. Thus, collaborative filtering makes predictions (i.e., filtering) about the interests of a user by collecting preferences from many users (i.e., collaborating). Collaborative filtering has been used in similar contexts [15] and location prediction [9].

Data Collection

Our data collection site was a major international airline with more than 165 destinations in dozens of countries. In the normal course of business, the organization sends marketing and advertising messages to customers in its frequent flyer database for various offers. The company also in the normal course of business performs various analysis to determine what marketing messages to send to which customers.

For this study, we partnered with the airline company to conduct the data analysis and then perform an A/B test in order to evaluate our research objectives. Using a data set of airline passenger flight bookings over a more than two-year period (2016-2018, we applied both the ensemble method and collaborative filtering method to identify customers and their next likely destinations. For each customer, we select the top five destinations that our algorithms predicted the customer would book next or like to book based on their historical booking pattern, i.e., precision at 5 (P@5). For performance comparison, we used a baseline of recommending the top most popular destination (i.e., using no algorithm but just recommending the most popular destination). The baseline has an accuracy P@5 of 19%. After training our models, we then tested on a set of approximately 100,000 customers. Based solely on historical data, the ensemble approach had an accuracy of 47.6% (148% better than baseline), and the collaborative filtering approach had an accuracy of 24.8% (30% better than the baseline).

From the complete set of frequent flyers, we randomly selected 10,000 customers for an actual user test. We split these 10,000 customers into 4,500 for the ensemble recommendation, 4,500 for the collaborative filtering recommendation, and 1,000 for a control group. The test groups were sent marketing messages crafted by the airline company marketing department and the control group was sent no marketing messages. Each marketing message contained a recommendation for one of the selected destinations and offered the customer an incentive to book. The incentive offered was bonus miles for a flight booking to one of the destinations within the offer window.

The user marketing messages (see Figures 2 and 3) were sent over a one-week period in 2018, with the participants having two-weeks to respond to the offer. The marketing message sent to each customer was identical except for the five recommended destinations. Each message was individually sent via an opt-in email promotional offer. The messages were crafted by the organization's marketing department and were personalized with the member's name and membership number. The customer could directly book the flight from the marketing message. In the end, 7,784 customers (generally 50%/50% between the two

Group	*Response*
Control	10.4%
Test	12.1%
Uplift	16.4%

Table 2: Percentage of bookings for the test groups and the control group, along with the uplift in bookings for the test group.

Method	*Results*
Ensemble	12.33%
CF	11.88%
Overall	12.10%

Table 3: Percentage of bookings for the ensemble and collaborative filtering test groups, along with overall response rate for the test groups. Accuracy is defined as the percentage of correct predictions.

models) actually received the offers, with the dropout rate being due to outdated email addresses and limitations on contacting customers within a given period. This highlights the practical issues of conducting user testing on real systems with real customers.

Results

Returning to our research objective (c), our combined algorithmic approaches resulted in prediction accuracy during the user testing of approximately 23% as measured by user selecting the predicted destination. The results of the A/B test showed a 16% increase in bookings of the test groups compared to the control group. The overall predictive power was 23%, with collaborative filtering having a predictive power of 30% and 19% for the ensemble method. Table 2 shows the overall results. Table 3 shows the results for each algorithmic approach and overall for both algorithms.

Discussion and Implications

The overall combined accuracy of the user test was 23%, substantially lower than the results based solely on historical data. While this is still 21.1% above the baseline of 19%, it is substantially lower than the accuracy using historical data. Compared to results using solely historical data, the ensemble approach was only 39.91% as accurate with actual customers; collaborative filtering approach performed 20.97% better with real customers versus historical data.

So, in both cases, using historical data did not represent the true predictive accuracy of the algorithms. The underlying tastes of the population may be flux and not reflected in historical data. From an analysis of the booking results, the customers who accepted the collaborative filtering offers booked at a higher rate. Our premise is that serendipitous nature of the novel destinations was enticing, which may have induced customers to book at higher rates than the ones suggested for the customers strictly on the predictive aspects, which shows that these approaches should be evaluated with real people.

Conclusion and Future Work

Results have implications in a variety of areas. We suggest in these contexts that organizations not rely on historical data as the users may change. The findings of this exploratory research are exciting as the foundation for future and fruitful, including segmenting the overall population via various demographic features, such as gender, age, and nationality, to evaluate whether or not these factors play into the context and also measuring the effect of the marketing messages.

Acknowledgements

We thank the international company for their collaboration for this research.

References

[1] J. Beel, M. Genzmehr, S. Langer, A. Nürnberger, and B. Gipp, "A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation," in *Proceedings of the international workshop on reproducibility and replication in recommender systems evaluation*, 2013, pp. 7-14.

[2] S. Chen, W. Huang, M. Chen, J. Zhong, and J. Cheng, "Airlines Content Recommendations Based on Passengers' Choice Using Bayesian Belief Networks," in *Bayesian Inference*, J. P. Tejedor, Ed., ed: IntechOpen, 2017.

[3] H. Fani, E. Jiang, E. Bagheri, F. Al-Obeidat, W. Du, and M. Kargar, "User community detection via embedding of social network structure and temporal content," *Information Processing & Management,* vol. 57, p. 102056, 2020.

[4] S. Fitchett and A. Cockburn, "AccessRank: predicting what users will do next," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, Texas, USA, 2012.

[5] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems,* vol. 22, pp. 5-53, 2004.

[6] C. Hueglin and F. Vannotti, "Data mining techniques to improve forecast accuracy in airline business," presented at the Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, CA, 2001.

[7] B. P. Knijnenburg and M. C. Willemsen, "Evaluating recommender systems with user experiments," in *Recommender Systems Handbook*, ed Boston, MA: Springer, 2015, pp. 309-352.

[8] R. D. Lawrence, S. J. Hong, and J. Cherrier, "Passenger-based predictive modeling of airline no-show rates," presented at the Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C., 2003.

[9] D. Lian, V. W. Zheng, and X. Xie, "Collaborative filtering meets next check-in location prediction," presented at the Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 2013.

[10] X. Ling, W. Deng, C. Gu, H. Zhou, C. Li, and F. Sun, "Model Ensemble for Click Prediction in Bing Search Ads," presented at the Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 2017.

[11] P. Pu, L. Chen, and R. Hu, "A user-centric evaluation framework for recommender systems," in *Proceedings of the fifth ACM conference on Recommender Systems*, 2011, pp. 157-164.

[12] S. J. Racine and J. P. Curtin, "Developing an airline freight management system: meeting airline and end-user challenges," presented at the CHI '03 Extended Abstracts on Human Factors in Computing Systems, Ft. Lauderdale, Florida, USA, 2003.

[13] S. Renjith, A. Sreekumar, and M. Jathavedan, "An extensive study on the evolution of context-aware personalized travel recommender systems," *Information Processing & Management,* vol. 57, p. 102078, 2020.

[14] P. Sánchez and A. Bellogíns, "Building user profiles based on sequences for content and collaborative filtering," *Information Processing & Management,* vol. 56, pp. 192-211, 2019.

[15] O. S. Shalom, N. Koenigstein, U. Paquet, and H. P. Vanchinathan, "Beyond Collaborative Filtering: The List Recommendation Problem," presented at the Proceedings of the 25th International Conference on World Wide Web, Montreal, Quebec, Canada, 2016.

[16] Y. Wang, C. Breitinger, B. Sommer, F. Schreiber, and H. Reiterer, "Comparing Sequential and Temporal Patterns from Human Mobility Data for Next-Place Prediction," presented at the Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, Singapore, Singapore, 2018.