Analyzing Complaints and Customer Satisfaction in the Travel Industry

Youssef Drissi, Markus Ettl, Anna Lisa Gentile, Scott McFaddin, Petar Ristoski, Wei Sun
IBM Research
{youssefd, msettl, annalisa.gentile, mcfaddin,sunw}@us.ibm.com, petar.ristoski88@gmail.com

Abstract

Customer satisfaction is crucial for the long term success of any travel service provider. Therefore, identifying situations that can lead to customer dissatisfaction is critical. The strongest evidence of customers dissatisfaction are their complaints. While complaints do not occur very often, they almost always lead to loss of customer goodwill which can cost travel providers millions of dollars in future revenues. In this paper, we describe an approach to proactively identify customers that have the highest propensity to complain as they encounter a travel disruption event. These are invaluable insights that can empower customer service teams with information to deliver a more timely, relevant and impactful service experience. We use three key aspects in this approach: (i) specialized feature engineering for the travel industry; (ii) handling extremely imbalanced data and (iii) adaptation of binary classification, anomaly detection and learning to rank models to our specific task. This research is an important step towards more individualized understanding of customer behavior, and potential service enhancements to further increase customer satisfaction.

1. Introduction

During a typical year there are more than 40 million commercial flights worldwide, carrying more than 5 billion customers.[1] Airlines around the world are competing for passengers' demand which makes the airline industry extremely competitive. From a customer's perspective, there often is very little distinction in travel choices between a given origin and destination. Similar flights are offered by multiple carriers, in many cases departing and arriving within minutes of each other. In addition, the difference in pricing between companies is often minimal. Therefore airlines increasingly strive to provide differentiation in customer service to ensure a superior customer experience.

While every flight or travel moment cannot be personalized, acknowledging people as individuals and customized messaging to customers is an easy way to let passengers know that they are seen beyond a price point. In particular, personalizing the experience for customers who are confronted with service disruptions such as arrival delays, flight cancellation or missed connections can avert a decline in the airline's Net Promoter Score (NPS)[2] and improve customer loyalty.

We worked with a major legacy airline on a proof of concept to proactively identify customers that have the highest propensity to complain as they encounter a flight disruption, thereby providing customer-facing teams with information to deliver a more timely, relevant and impactful service experience. The airline operates as the dominant carrier at several major hub airports and fields a substantial staff of customer service representatives to address customer issues. Ideally the airline would like to dedicate staff to respond to customers who have experienced flight disruptions on the spot. In order to deploy the available customer service staff most effectively, the airline sought a solution enabling a two step response: (1) identify flights with the highest propensity to generate complaints and dispatch representatives to greet those flights, and (2) identify customers on those flights with the highest need of attention. This two-step approach is reflected in our solution described below.

The task of learning patterns that lead to a complaint from historical booking and flight data is challenging because the occurrence of complaints is relatively rare compared to the total passenger count. From a machine learning point of view, the available data is highly imbalanced, where complaints represent less than 1% of all the data. In this work we tackle this issue by splitting the problem into two parts: at first we identify

[1] Data obtained from https://www.statista.com/statistics/564769/airline-industry-number-of-flights/

[2] NPS is a metric used in customer experience programs to measure the loyalty of customers to a company

URI: https://hdl.handle.net/10125/79567
978-0-9981331-5-7

flights which are most likely to generate a customer service issue, and then we learn a ranking for passengers on those flights where the top-scoring travelers are predicted as most likely to complain. In each of these two steps we introduce a number of technical novelties.

The major contributions of this work are threefold. First and foremost, we design a feature engineering pipeline which is specific for the travel industry. To the best of our knowledge, there are no publicly available guidelines to prepare such data. We then design the problem as a "divide and conquer strategy" to handle imbalanced data, and experiment with state of the art machine learning models - including neural models and ranking models - for the specific task. Lastly, we adapt an existing learning to rank model, typically used in information retrieval, for the task of ranking customers on a flight based on their propensity to complain. We employ DBSCAN clustering to identify important customer clusters, which are then used for feature weighting in the learning to rank model. We evaluate the approaches on an actual flight data set from 2019, achieving precision@1% of 95.9% at the flight level, and precision@1 of 63.2% at the customer level. We would like to stress that, more than technical novelties, the main contribution of this work is the methodology for effectively solving a real business problem through careful combination and adaptation of the right state-of-the art algorithms.

We give an overview of related work in Section 2, and formally define the problem and describe our solution in Section 3. We present our results in an offline test in Section 4 and discuss the pilot implementation in Section 5. Lastly, we draw conclusions and discuss potential future work in Section 6.

2. State of the art

In the highly competitive airline industry, the importance of customer satisfaction is paramount for customer engagement and retention [1], and travelers' perception of service quality has been the subject of studies for years. Gan et al. [2] identified seven dimensions which are positively related to perceived service quality: timeliness, assurance, convenience, helpfulness, comfort, meals, and safety. The perception will differ according to passengers' age, gender, income, occupation and marital status. According to [3, 4] perception will also differ between business and leisure travelers, where demographic variables such as gender, income and education are statistically significant for one group of passengers but not for another. Similarly, Climis et al. [5] concluded that models for customer retention are affected depending on which group the travelers belongs to, according to the purpose of their travel: business, education, vacation or family visit. We ground our work on these findings and pay close attention to passenger features when developing customer-level models.

Regardless of how well an airline is doing, there will always be a percentage of customers who complain. Complaints can be a tremendous source for learning pain points and improving the business and ultimately avoid or minimize customer churn, i.e. the loss of existing customers to a competitor. Many studies tried to analyse the causes of complaints for the airline industry. Chow et al. [6] analyzed customer complaints from twelve large and small Chinese airline carriers, and found that on-time performance indeed plays a role in customer complaints depending on the difference between actual and expected on-time performance. Passengers' loyalty status plays a role in customers' expectations about the handling of a complaint [7]: higher loyalty tier customers were found to be more likely to expect airline personnel to comply with their demands, even when demands are unreasonable. Nonetheless, correctly handling complaints will increase customer satisfaction and customer engagement [8].

Following these guidelines, we focus our work on proactively identifying customers who are more likely to complain when problems arise. These insights are invaluable when devising customer care policies that implement proactive actions to maximise travellers' satisfaction. To the best of our knowledge there are no known models in the literature that predict the propensity of a customer to complain, especially in the travel industry. On the other hand, there is an abundance of classification models proposed for churn prediction, including: Support Vector Machines, Naïve Bayes, Decision Trees and Neural Networks [9]; Support Vector Data Description (SVDD) with random under-sampling and SMOTE oversampling [10]; combinations of random under-sampling and boosting algorithm [11]; random forest combined with random oversampling [12]; Multilayer Perceptron (MLP) neural network [13]; Reverse Nearest Neighborhood and One Class support vector machine (OCSVM) [14]; hybrid combination of well known oversampling technique SMOTE with under-sampling technique [15]; ensemble learning [16] and transfer learning methods [17]. Both complaints and churn are relatively rare events, and building statistical patterns to predict them is extremely difficult due to the imbalance of the data sets: one class (the complaints/churn) is much smaller than the other classes. Recent methods apply machine learning techniques and address the class-imbalance problem with over/under sampling techniques [18, 19].

The rate of customer complaints in the data set used in this study is less than 0.3%, which makes the data set extremely imbalanced. In a pre-study we found that standard over sampling of the minority class alone fails to achieve satisfactory classification results. Therefore, we look at the class imbalance problem from a different angle and propose a divide and conquer strategy. First, we identify which flights will likely generate some complaints. Second, we model each customer propensity to complain within the scope of each flight as a ranking problem, which significantly reduces the problem of dealing with class imbalances (with the trade-off of potentially missing customers on regular flights who have a high propensity to complain). In terms of predicting flights which are likely to generate some complaints we rely on binary classification models. In terms of predicting the customer propensity to complain, we model it as a ranking problem using learning to rank models with feature weighting. To the best of our knowledge it is the first time that customer propensity to complain is modeled locally and addressed as a ranking problem.

3. Method

The input of the system is a set of flights $F = \{f_1, f_2, ..., f_m\}$, where each flight is represented with features $X_f = \{x_{f1}, x_{f2}, ..., x_{fk}\}$ of size k. On each flight f there is a set of n customers $C_f = \{c_1, c_2, ..., c_n\}$, where each customer is represented with features $X_c = \{x_{c1}, x_{c2}, ..., x_{cl}\}$ of size l.

Our approach follows a two-step method: (i) given a list of flights F_t in a time period t it ranks the flights in descending order based on the probability of a given threshold number of customers on the flight to complain; (ii) given a flight f and the list of customers on the flight $C_f = \{c_1, c_2, ..., c_n\}$ the approach ranks the customers in descending order based the probability to complain, such that the probability to complain of the first ranked customer is higher than the probability to complain of the second customer $P(c_1|f) > P(c_2|f)$, or in general $P(c_{n-1}|f) > P(c_n|f)$ and $P(c_1|f) >> P(c_n|f)$.

A flow diagram of the overall methodology is shown in Figure 1. In the first module of the system, the data is pre-processed and flight and customer features are extracted. In the second module, we develop a flight-based model, which ranks the input flights based on the probability for a given threshold number of customers on a flight to complain. The third module receives a list of flights from the previous module, and ranks the customers on each flight based on the probability to complain. The machine-learned ranking model aims to produce an optimal ranking of customers per flight relative to their probability to complain, using different customer segments that are derived from booking and loyalty features. The output of the system is a ranked list of passengers per flight, based on a ”customer dissatisfaction score”, normalized between 0 and 100.

3.1. Flight-Level Model

To train a machine learning model for flight-level complaint prediction, we first extract a set of features for each flight instance $X_c = \{x_{f1}, x_{f2}, ..., x_{fl}\}$. We consider 5 types of flight level features for training and one type of feature for labelling:

- **Aggregated customer & loyalty program features** include an aggregated view of the customers on the flight such as the percentage mix of loyalty customers among the boarding passengers. Aggregated customer information can have a substantial differentiating effect among similarly delayed and otherwise operationally affected flights, for example, flights with higher degrees of loyalty customers tend to have fewer complaints.

- **Scheduling features** such as the scheduled and actual departure time, arrival times, duration, day-part, day-of-week, season as well as equipment capacities and fill-rates (“crowding”) have a substantial effect on complaint propensity. For example, flights scheduled for arrival later in the day can be seen to have higher complaint rates. Additionally, differentials in these values can be used to train a “surprise” factor.

- **Operational features** relate to the actual operational aspects of a flight, such as departure and arrival delays, delays in the air (e.g. re-routings due to weather), and taxi-in and taxi-out times. In some cases the flight is outright canceled. Additionally important are various measures of controllability. In many cases it is found that passengers are more sensitive to one type of operational problem but not to its symmetric counterpart (for example, delays in taxi-out time are found to be more important than delays in taxi-in time).We also found that complaint propensities are often mitigated by both qualitative and quantitative perceptions of controllability.

- **Historical flight complaint rates** are established

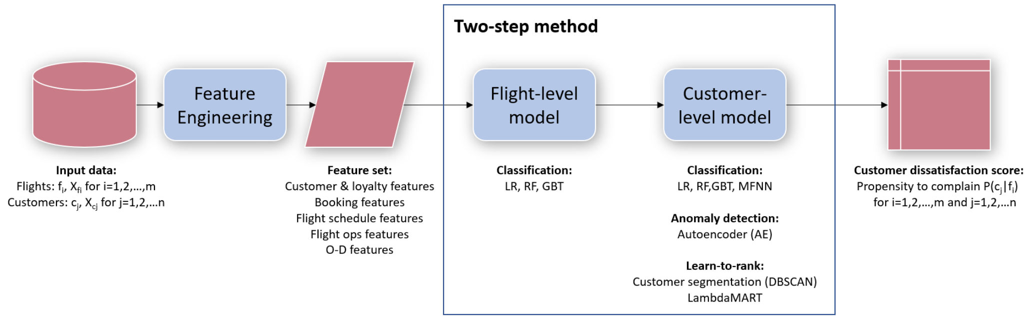


Figure 1: Flow diagram of the overall modeling approach.

by a reverse analysis of the training data set, keyed across various dimensions, e.g. by the flight number (e.g. "Flight 7"), origin-destination pairings (e.g. "LAX-SFO") or combinations of the route and time of departures (e.g. "LAX-SFO-Morning").

- **Origin and destination features** use aspects of the originating and arriving cities as a feature, as well as a measure of "hubiness" which is higher for certain interchange points and lower in cases of terminals used both for interchange and as a destination. Highly hub-like destinations, for example, place substantial connection pressure on passengers and can cause higher complaint rates.

The final set of features is a mixture of categorical and numerical features. The numerical features are standardized with mean 0 and standard deviation 1. The categorical features are converted to one-hot encoding representations.

We consider the task of flight-level complaint prediction as a binary classification task, for which we used Logistic Regression (LR), Random Forests of Decision Trees (RF), and Gradient Boosted Trees (GBT) model. To further improve the performance, we employed a "data lensing" approach in which we altered the conditions which determine the target variable of the classification method. With this technique, we redefine the definition of the positive case (the "Y variable") to be true only for instances of flights that produced 3 or more complaints rather than 1 or more complaints – this higher threshold was empirically determined. Note that this process only affects the internal training of the flight model and does not introduce a limitation of how the flight model can be applied.

3.2. Customer-Level Model

To train a machine learning model for customer ranking, we first extract a set of features for each customer $X_c = \{x_{c1}, x_{c2}, ..., x_{cl}\}$. In addition to the **Flight operations features** explained in the previous section, we consider two additional types of customer features. The **Customer & loyalty program features** include demographics such as age group or preferred language, and airline loyalty features such as membership status, lifetime flown miles, lifetime spent money, airline awards etc. These features are updated after each new flight. The **Booking features** are related to each booked flight by the customer and include: leg origin, leg destination, type of flight (domestic or international), booking channel, number of hops, ticket type, etc. as well as: advance purchase in days (which often correlates with type of travel, i.e. leisure or business), number of previous complaints, amount of compensation received on previous flights, number of disrupted flights in the past and number of travel companions (all good indicators of customer satisfaction). Concatenating the flight features to the customer features allows us to identify different patterns and combination of customer and flight features that might lead to increased propensity to complain. As with the flight-level model, we standardize the numerical features, and convert the categorical features to one-hot encoding representations.

Next, we model the task of predicting customer propensity to complain using 3 different approaches: (i) binary classification problem; (ii) anomaly detection and (iii) learning to rank.

3.2.1. Binary Classification We consider the task of ranking customers on a flight as a standard binary classification problem, where we use the classification model confidence score for ranking the customers on each flight. To address the extreme imbalance in the data set, we under-sample the majority (negative) class, and over-sample the minority class. We build four different binary classification models: Random Forest (RF), Logistic Regression (LR), Gradient Boosted Trees (GBT), Multilayer Feedforward Neural Network (MFNN).

3.2.2. Anomaly Detection In extremely imbalanced data sets, anomaly detection approaches can be used. In this case, we consider the negative class (customers that do not complain) as the "normal" data, while the positive class (customers that complain) is the outlier/anomaly. Such outlier approaches aim to model the distribution of the "normal" class during training time, and all instances that do not fall under the learned distribution will be considered as outliers, i.e., we expect that there are some irregularities and patterns in the feature vectors of the positive instances that make them differ from the most of the data in the data set. Autoencoder neural networks [20] can be used for anomaly detection, and have shown outstanding performance in the past [21]. An autoencoder is a feed forward neural network, which represents the state-of-the-art for unsupervised representation learning tasks. Autoencoders consists of two parts, an encoder and a decoder. The encoder takes the data on the input and tries to compress it to a much smaller vector representation, which retains only the most important features. The decoder learns how to reconstruct the original data from the compressed encoded representation, producing a representation of the data that is as close as possible to the original input data. The autoencoder is trained only on the majority (negative) class, in order to learn how to compress and re-create the "normal" data in the data set. By doing so, it is expected that the autoencoder won't be able to compress and recreate the positive class as good as the negative, i.e., the reconstruction error is expected to be high for the positive instances and therefore identify them as anomaly. We use the autoencoder reconstruction error to rank the customers on each flight, i.e., the higher the reconstruction error, the higher the probability to complain.

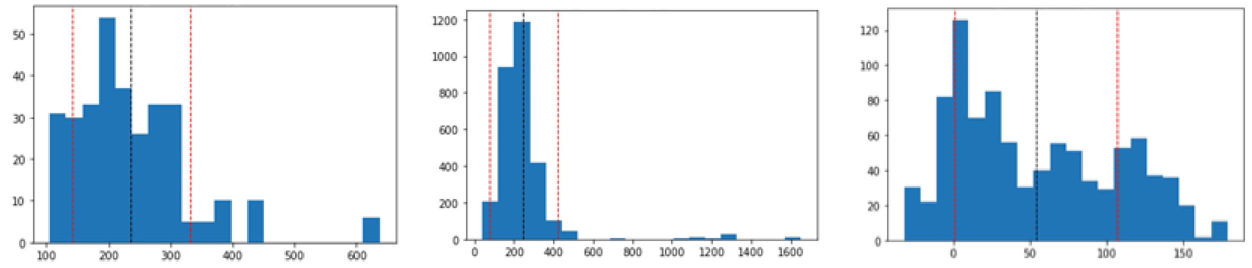
3.2.3. Learning to Rank Traditional ML models (e.g. binary classification models) build a single model, or a set of models, to make a prediction on a single customer at a time, by assigning a numeric score of the likelihood of the positive class. On the other hand, learning to rank models aim to produce an optimal ranking of customers per flight, based on their probability to complain. Instead of optimizing the numeric score for each customer, the model tries to optimize the rank of the complete list of customers, where the model tolerates fewer errors at higher ranked positions, i.e., the top N ranked customers are expected to be the customers with a higher probability to complain.

While there are many learning to rank algorithms in the literature, for this task we use the LambdaMART [22] ranking algorithm with boosted trees, which uses the pairwise-ranking approach to minimize the pairwise loss by sampling many pairs of instances in the data set.

To be able to use learning to rank models, first we have to define *groups*. We create a group for each origin-destination pair. For example, for the flight *Los Angeles to San Francisco* we create a group *LAX-SFO*. To train the model, we group the customers based on the flight groups, and assign the complaint score, i.e., 0 if they did not complain and 1 if they complained. Each customer is represented with a feature vector, as explained before.

One of the drawbacks of such an approach is that the flight-level features are exactly the same for all the customers on the same flight. However, the flight-level features can have significantly different effect on the customers. For example, on a delayed flight, customers traveling for leisure would have a higher propensity to complain than frequent fliers. To support this claim we perform clustering on the customers, based on the *Customer & Loyalty program* and *Booking* features. To do so, we use the density-based DBSCAN clustering algorithm [23], which clustered all the customers in 6 clusters. With manual analysis we were able to assign a descriptive label to each cluster, i.e., "Elite Customers", "Corporate Customers", "Corporate Elite Customers", "Leisure Single Customers", and "Leisure Couple/Family Customers".

Figure 2 shows the distribution for the flight arrival delay (in minutes), which differs significantly across the three clusters of customers. We integrate such information in the learning to rank model: (i) calculate the distribution for each flight-level feature for each of the customer clusters and calculate the mean and standard deviation; (ii) assign the customers on each flight to one of the existing customer clusters; (iii) for each customer in each group in the learning to rank model weight the flight-level feature based on the cluster to which the customer belongs, i.e., the feature value x_c is replaced with a weighted value x_{cw} equal to the number of standard deviations the value falls to the right or to the left of the mean of the feature distribution X_d in the given cluster C_c using Equation 1.



a) *Elite customers*

b) *Corporate elite customers*

c) *Leisure single customers*

Figure 2: Value distributions for the arrival delay feature for different customer clusters. The black dashed lines represent the mean, and the red dashed lines represent the standard deviation.

$$x_{cw} = \frac{x_c - mean(X_d)}{std(X_d)} \tag{1}$$

For example, if a flight is being delayed for 200 minutes, based on the distributions shown in Figure 2, the feature will have value -0.38, -0.28 and 3.08, for each cluster respectively. This means that such a delay will have a much bigger impact on the likelihood to complain for a leisure customer than for an elite customer.

4. Experiments

The data set used in our evaluation is based on twelve months of booking data from domestic and international markets served by the airline with over 280 million flown flight segments. The primary focus of this work is to study customer dissatisfaction caused by travel disruptions such as flight delays, cancellations or missed connecting flights. As such, the data set only contained customer complaint cases where the reason was categorized as “flight disruption”. Complaints caused by to other incidents (such as poor onboard experience, airline staff behavior, or information handling issues) were not considered. The data set contains over 500,000 flight segments that were delayed or cancelled, out of which 9.26% flights had at least one customer complaint. The total number of customers on all flights in the data set is more than 35 million customers, of which only 0.269% customers complained. This makes the data set extremely imbalanced. We use this data set to evaluate both the flight-level model and customer-level model.

4.1. Flight-Level Model

To evaluate the flight-level model we use two sets of metrics. First, considering the problem as a standard binary classification problem, to evaluate the models we use Precision (P), Recall (R) and F-score (F) on the positive class. Second, considering it as a ranking problem, i.e., ranking the flights based on the propensity to complain per day, we use precision@N and recall@N, calculated as the average precision@N and recall@N per day. The average number of flights per day is over 2,000, therefore we report precision and recall at top 1%, 3% and 5%, addressing a rate of attention of approximately 100 flights per day.

The results are calculated using stratified 10-fold cross validation. We make sure that all the flights in a single day are either in the training or test set exclusively. The final results for each model are shown in Table 1.

To calculate the F-score, we perform threshold moving used to map probabilities to class labels. In each validation fold, we use a hold out data sets to identify the optimal probability threshold by analyzing the precision-recall curve, i.e., we choose the threshold that yields the best trade-off between precision and recall on the hold out data set. From the results we can observe that the GBT model significantly outperforms the rest. We note that based on the real application of our system, achieving high precision is crucial compared to recall, as there is only a constrained number of flights that can be recommended for attention. Therefore the model needs to output highly precise recommendations at the top-most positions.

4.2. Customer-Level Model

We evaluate all three types of customer-level models, i.e., binary classification models, autoencoders for anomaly detection (AE) and learning to rank model (LTR). To train the binary models, we first re-sample the data set, i.e., we over-sample the minority class by a factor of 5, and we under-sample the majority class to have the same size as the minority class after over-sampling. The multilayer feedforward neural network consists of an input layer, and 4 dense layers with size 200, 100, 50 and 30, respectively, using ReLU activation function. The output is calculated using a

Table 1: Results for the flight-level model for boosted trees binary classification models using data lensing.

Method	P	R	F	P@1%	P@3%	P@5%	R@1%	R@3%	R@5%
LR	48.5	52.2	49.6	85.0	80.8	76.3	4.4	11.7	18.3
RF	49.8	57.7	52.9	92.1	83.9	79.1	5.1	12.3	18.5
GBT	**52.4**	**59.4**	**55.2**	**95.9**	**88.9**	**83.7**	**5.0**	**13.0**	**19.7**

Table 2: Results for the customer-level model for binary classification models, autoencoder and learning to rank model.

Method	P	R	F	P@1	P@3	P@5	R@1	R@5	R@10
RF	9.10	13.74	10.94	31.21	19.04	13.75	31.21	53.75	62.16
LR	6.09	15.64	8.76	27.55	17.36	13.26	27.55	50.67	59.98
GBT	11.71	15.76	13.43	32.78	19.62	14.93	32.78	54.21	64.81
MFNN	8.64	31.67	**13.58**	29.28	13.44	9.63	29.28	33.18	47.30
AE	7.90	14.48	10.23	32.43	13.98	12.32	32.43	54.46	59.82
LTR	/	/	/	**63.18**	**29.93**	**20.85**	**63.21**	**79.88**	**85.25**

softmax layer. The architecture of the autoencoder is as follows: the encoder consist of an input layer with a size of the number of features, 4 dense layers with ReLU activation function with 200, 100, 50 and 40 units in each layer, respectively; the decoder consist of 4 dense layers with ReLU activation function with 50, 100 and 200 units in the first three layers, and the last layer has the same size as the input layer in the encoder. For the learning to rank model we tune the tree-based parameters and the regularization parameters to achieve best performance. To implement the binary classification models, we use the scikit-learn library[3]; to implement the neural network and the autoencoder, we use the Keras API[4]; to implement the learning to rank model we use the XGBoost library[5].

To evaluate the models we use two sets of performance metrics. First, considering the problem as a standard binary classification problem, to evaluate the models we use Precision (P), Recall (R) and F-score (F) on the positive class. For these metrics we use the complete data set, i.e., including flights on which no customer complained. Second, considering it as a ranking problem, the metrics that we use to evaluate the model are: Precision@N, which is the average Precision@N of all the flights, where precision@N on a single flight is the fraction of correctly identified customers that complained in the top N ranked passengers; Recall@N, which is the average recall@N of all the flights, where recall@N is the fraction of all the customers that complained identified by the model. Calculating precision@N and recall@N for flights on which no customer complained will always yield 0, therefore, for these metrics we use only the flights on which at least one passenger complained, i.e., we assume perfect performance from the flight-level model in identifying all the flights with at least 1 complaint. This data set consists of more than 50,000 flights with around 4 million passengers, of which only 2.26% passengers complained. The average number of passengers per flight is over 65, therefore we report precision at 1, 3 and 5, and recall at 1, 5 and 10.

For the binary classification models and the autoencoder we calculate both metrics, while for the learning to rank model only the second set of metrics. The results are calculated using stratified 10-fold cross validation. We make sure that all the passengers on the same flight are either in the training or test set exclusively. The final results for each model are shown in Table 2.

For the binary classification models and the autoencoder we perform threshold moving used to map probabilities to class labels, as done for the flight-level model.The results show that the multilayer neural network achieves the best results among all binary classification models and the autoencoder, when considering the task as a binary classification task. However, when we consider it as a ranking task, the learning to rank model significantly outperforms all the other approaches. The gradient boosted trees model shows promising results, followed by the autoecoder.

Besides achieving high performance, for such applications it is of great value to provide interpretability of the model predictions. To do so, we use SHAP (SHapley Additive exPlanations) analysis [24]. SHAP analysis can be used for global and local interpretability. In global interpretability, SHAP values indicate how much each feature contributes, either positively or negatively, to the target variable while in local interpretability, the SHAP values indicate how each

[3] `https://scikit-learn.org/`

[4] `https://keras.io/`

[5] `https://xgboost.readthedocs.io/en/latest/`

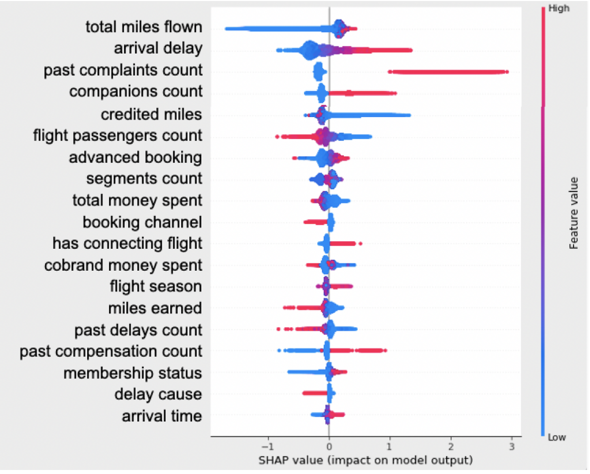


Figure 3: SHAP analysis for the L2R model.

feature contributed for the model prediction for a single instance.

Figure 3 shows the SHAP value plot for the L2R model. The plot shows the positive and negative relationships and importance of the features with the target variable. For example, the total miles flown by a customer, the arrival delay, the number of complaints in the past and the number of travel companions have a high impact on the model and are positively correlated with the target variable. In Section 5 we show how SHAP analyses are used to provide local interpretability, and how they can help customer care agents with their handling of customer complaints.

5. Solution Architecture and Customer Care Insights

Next we describe a hybrid cloud platform that was developed in this proof of concept. Figure 4 shows the integration architecture of the hybrid cloud solution. The platform enables the collaborative building of AI models on large customer data sets and the hosting of analytical solutions that leverage services and patterns offered by the platform. A web-based dashboard allows customer care agents to search flights (or passengers) and visualize the results of the underlying propensity models. For each customer, the dashboard provides a customer dissatisfaction score in the context of a specific journey, i.e., the propensity to complain resulting from the customer's travel experience on that journey. It also displays flight-level complaint scores that represent the estimated propensity of a flight to have one or more passenger complaints. In addition, the dashboard views illustrate the main drivers of customer dissatisfaction as derived from the local interpretability model described in section 4.2.

The raw data is transferred from a data lake hosted in the airline's IT environment into the landing zone of a Big Data analytics cloud platform where the raw data is processed to obtain basic flight and passenger features. Additional calculated features are derived from the data in the raw and unified zones and stored as feature vectors for machine learning. A training data set is prepared and used to train the flight propensity model, and another data set is prepared and used to train the passenger propensity model. The results from each model together with the SHAP analysis results are stored in the insight zone.

The web-based dashboard consists of five main components. The *Summary View* serves as a landing page and provides a compacted view of historical data and model prediction results. The *Flight Search* and *Passenger Search* components allow the user to search for flights or passengers using various search parameters. Common search parameters in Flight Search include flight origin airport code, flight destination airport code, date range, operating carrier, and flight number. Once a flight record is retrieved, the user interface shows a *Flight Dashboard* which includes a summary of flight data (such as flight number, number of enplaned passengers, and delay statistics), the flight complaint propensity score, and a list of booking records for passengers boarded on that flight. Passenger Search allows to query a passenger's flight activity using a unique passenger identifier or a record locator.

Once a passenger record is selected, the user interface displays the *Passenger Dashboard* shown in Figure 5 which contains a summary view of the passenger data in the context of a selected flight segment, the passenger complaint propensity score, past flight disruption records, and a chart that illustrates the top-scoring features with positive and negative impact on the customer's complaint propensity based on their SHAP values. Features with positive SHAP values are shown with a horizontal bar chart oriented to the right side, and features with negative SHAP values are shown with a bar chart oriented to the left side. The length of each bar represents the magnitude of the corresponding SHAP value.

The customer shown in Figure 5 is travelling on a domestic flight that has an arrival delay of 70 minutes. The predicted dissatisfaction score of 91 is among the highest scores of all passengers that boarded this flight. Features that contribute positively to the customer's dissatisfaction score ("aggravation effects") include: past complaint activity (the customer complained on three previous occasions and received compensation), experience of a moderate delay, and travel on an inbound

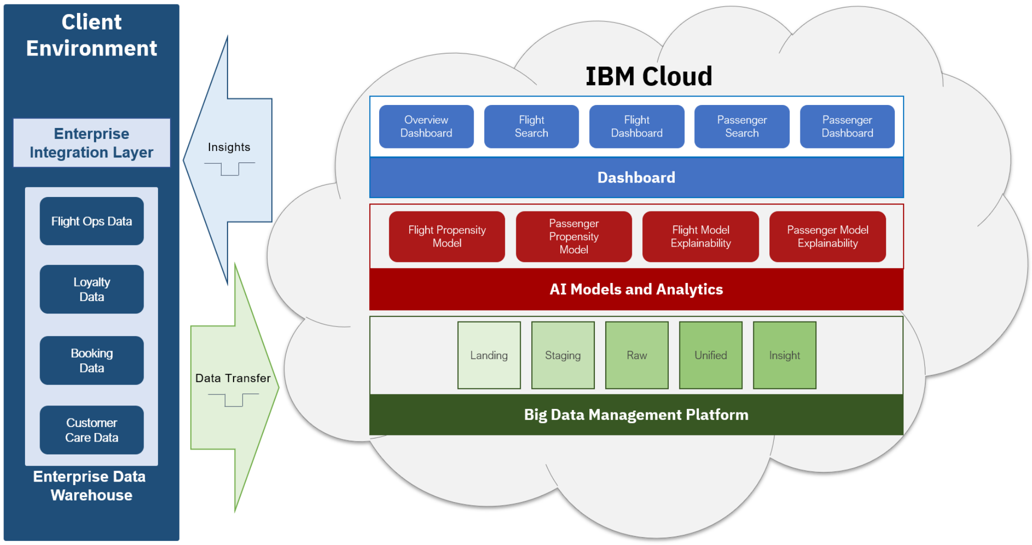


Figure 4: Hybrid cloud solution architecture.

connecting flight. Factors that contribute negatively to the propensity score ("mitigation effects") include: elite frequent traveler status, short advance booking window and travel without a companion (the customer is a high-tier loyalty program member, late booker, and traveling alone which are features typically associated with business travellers).

6. Conclusions and Future Work

Understanding customer dissatisfaction is important because customers are more likely to remember negative experiences as opposed to the positive encounters they've had with a service provider. Therefore an ability to anticipate when a customer is likely to complain about a service experience plays an important role in delivering a personalized experience and increasing customer loyalty and retention. This work focuses on the airline industry. We described a novel method to predict the likelihood of a traveler to complain and showed how we tested it. More specifically, we proposed a two-step approach where we first identify flights with an elevated risk of a service disruption, and subsequently rank passengers on each flight according to their propensity to complain. We validated the approach in a proof of concept with a global legacy airline, and performed a formal evaluation of the method on a large-scale travel data set. The results revealed far superior prediction results of our method when compared to conventional approaches based on classification models and auto-encoders.

The model insights can be an effective means for proactive customer engagement, rather than simply reacting to customer complaint issues. Personalized messaging in the context of a travel disruption, or proactively compensating a customer as a situation warrants, shows that the airline is customer-focused, and always striving to address issues as promptly as possible. This maximizes customer loyalty and increases the long-term value of customers as they continue to engage. In addition, insights from the customer-level models could be utilized to more efficaciously optimize compensation for disrupted customers during pre-travel (off-boarding in oversold situations) or post-travel (flight delays or cancellations) stages of a customer journey.

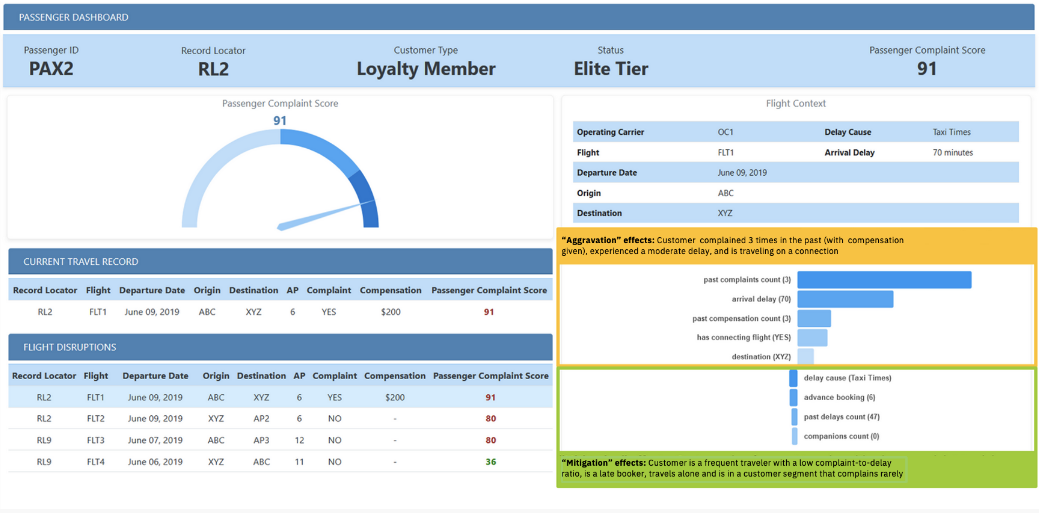
In future extensions of this work, we will investigate the use of time-series models to better assess each traveler and as a result create an even more personalized user experience and more customized frontline services. Most customers understand that things can and will go wrong. Making sure that customer care teams have all the relevant insights to provide apt resolution for customer grievances will only foster customer lifetime value and loyalty.

References

[1] R. Hapsari, M. D. Clemes, and D. Dean, "The impact of service quality, customer engagement and selected marketing constructs on airline passenger loyalty," *International Journal of Quality and Service Sciences*, 2017.

[2] C. Gan, "An empirical analysis of customer satisfaction in international air travel," *Innovative Marketing*, 2008.

[3] H. Jiang and Y. Zhang, "An investigation of service quality, customer satisfaction and loyalty in China's



PASSENGER DASHBOARD

Passenger ID	Record Locator	Customer Type	Status	Passenger Complaint Score
PAX2	RL2	Loyalty Member	Elite Tier	91

Flight Context

Operating Carrier	OC1	Delay Cause	Taxi Times
Flight	FLT1	Arrival Delay	70 minutes
Departure Date	June 09, 2019		
Origin	ABC		
Destination	XYZ		

CURRENT TRAVEL RECORD

Record Locator	Flight	Departure Date	Origin	Destination	AP	Complaint	Compensation	Passenger Complaint Score
RL2	FLT1	June 09, 2019	ABC	XYZ	6	YES	$200	91

FLIGHT DISRUPTIONS

Record Locator	Flight	Departure Date	Origin	Destination	AP	Complaint	Compensation	Passenger Complaint Score
RL2	FLT1	June 09, 2019	ABC	XYZ	6	YES	$200	91
RL2	FLT2	June 09, 2019	XYZ	AP2	6	NO	-	80
RL9	FLT3	June 07, 2019	ABC	AP3	12	NO	-	80
RL9	FLT4	June 06, 2019	XYZ	ABC	11	NO	-	36

"Aggravation" effects: Customer complained 3 times in the past (with compensation given), experienced a moderate delay, and is traveling on a connection

"Mitigation" effects: Customer is a frequent traveler with a low complaint-to-delay ratio, is a late booker, travels alone and is in a customer segment that complains rarely

Figure 5: Passenger Dashboard.

airline market," *Journal of Air Transport Management*, 2016.

[4] C. M. Ringle, M. Sarstedt, and L. Zimmermann, "Customer satisfaction with commercial airlines: The role of perceived safety and purpose of travel," *Journal of Marketing Theory and Practice*, 2011.

[5] R. Climis, "Factors affecting customer retention in the airline industry," *Journal of Management and Business Administration. Central Europe*, 2016.

[6] C. K. W. Chow, "On-time performance, passenger expectations and satisfaction in the Chinese airline industry," *Journal of Air Transport Management*, 2015.

[7] W. B. Chtou, M. H. Chang, and C. C. Yhng, "Customers' expectations of complaint handling by airline service: Privilege status and reasonability of demands from a social learning perspective," *Psychological Reports*, 2009.

[8] J. Cambra-Fierro, I. Melero-Polo, and F. Javier Sese, "Can complaint-handling efforts promote customer engagement?," *Service Business*, 2016.

[9] K. Eria and B. P. Marikannan, "Systematic Review of Customer Churn Prediction in the Telecom Sector," *Journal of Applied Technology and Innovation*, 2018.

[10] S. Maldonado, "Churn prediction via support vector classification: An empirical comparison," *Intelligent Data Analysis*, 2015.

[11] E. Dwiyanti, Adiwijaya, and A. Ardiyanti, "Handling imbalanced data in churn prediction using RUSBoost and feature selection (Case study: PT. Telekomunikasi Indonesia regional 7)," in *Advances in Intelligent Systems and Computing*, 2017.

[12] A. Hanif and N. Azhar, "Resolving Class Imbalance and Feature Selection in Customer Churn Dataset," in *Frontiers of Information Technology*, 2018.

[13] M. R. Ismail, M. K. Awang, M. N. A. Rahman, and M. Makhtar, "A multi-layer perceptron approach for customer churn prediction," *International Journal of Multimedia and Ubiquitous Engineering*, 2015.

[14] G. G. Sundarkumar and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance," *EAAI*, 2015.

[15] U. R. Salunkhe and S. N. Mali, "A hybrid approach for class imbalance problem in customer churn prediction: A novel extension to under-sampling," *International Journal of Intelligent Systems and Applications*, 2018.

[16] J. Xiao, L. Xie, C. He, and X. Jiang, "Dynamic classifier ensemble model for customer classification with imbalanced class distribution," *Expert Systems*, 2012.

[17] J. Xiao, Y. Wang, and S. Wang, "A dynamic transfer ensemble model for customer churn prediction," in *Business Intelligence and Financial Engineering*, 2014.

[18] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, "Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study," *IEEE Access*, 2016.

[19] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, 2019.

[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," tech. rep., California Univ San Diego, 1985.

[21] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *International Conference on Knowledge Discovery and Data Mining*, pp. 665–674, 2017.

[22] C. J. C. Burges, "From RankNet to LambdaRank to LambdaMART: An Overview," Tech. Rep. MSR-TR-2010-82, jun 2010.

[23] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.

[24] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *ArXiv*, vol. abs/1705.07874, 2017.