Machine Learning I

Hotel California

Predictive Overbooking Models

Julius Kasiske

23.12.2022

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

ABSTRACT

This report entails a detailed analysis of the cancellation behaviour of guests of Hotel California in Lisbon in 2016. It provides hotel management with insights into how they can reduce cancellations in the first place, but additionally entails a predictive model that detects cancellations before they occur at a performance of 0.79234 out of 1 (with 1 being the best). The performance was measured using the F1-score, a common evaluation metric for classification problems like ours. Our analysis shows that hotel management may be able to reduce cancellation rates of their guests by decentivizing online bookings with long lead times. They can manage these incentives through optimized pricing models, carefully curated refund policies and a comprehensive overbooking strategy. Aligning these three elements will restore relative profit-optimality that is currently being deviated from due to a cancellation rate of 37% without strategic responses to it.

KEYWORDS

Predictive Overbooking; Supervised Learning; Predictive Models; Classification

INTRODUCTION

When booking a hotel or any overnight stay, there is a distinct possibility for unexpected circumstances that yield the necessity to cancel the booking that was made. Guest may cancel due to sickness, family circumstances or additional reasons. Independent of their respective refund policy, hotels are then left with vacant rooms. These cancellations are often spontaneous, causing hotels to have to scramble to get the newly vacant rooms booked on short notice. Hotels then have to adjust prices, causing deviations from profit-optimality.

Thus, if hotels could predict which customers will cancel their booking, they would be able to strategically overbook rooms at prices that partially restore said profit-optimality.

With that potential in mind, we were tasked with building predictive models for the Hotel California and thereby support their attempt to formulate a novel overbooking strategy. Formulating such a strategy additionally requires a qualitative understanding of the determinants behind cancellations on the part of hotel management. We set out to provide that insight to hotel management, delivering actionable recommendations that enable the formulation of a successful overbooking strategy.

The factors driving cancellation probability for a given guest are well studied. We recognize the level of insight that research done by others can provide to Hotel California and thus based our exploratory analysis on findings that Martin Falk and Markku Vieru outlined in their paper "Modelling the Cancellation Behaviour of Hotel Guests" in January 18th, 2018[1], by validating subsets of their hypotheses and findings on our data. Thus, we aim to confirm that general patterns that were prevalent elsewere also pertain to the data provided by Hotel California. However, we simultaneously acknowledge the breath of possible determinants that are contained in the data set from Hotel California and consequentially explored additional explanatory variables that Falk and Vieru have not looked at.

The subsequent findings, while providing standalone value through the qualitative picture they help to paint, also help to contextualize the results of machine learning models. After all, cancellations will be predicted using these quantitative models.

HYPOTHESES

Hypothesis 1: Booking and cancellation behavior of guests is not cyclical
Hotel California is located in Lisbon and targets business travellers (all reservations in 2016 were made by companies). Business travel frequence usually decreases at the beginning and the end of each year due to the holiday season, but remains relatively stable throughout the remainder of each year, making it presumably less cyclical and volatile than leisure travel. Relative stability in booking frequency simplifies further analyses, since additionally analyzed relationships do not have to be stratified by season.

Hypothesis 2: Falk and Vieru's key findings are observable on Hotel California's data
Falk and Vieru found that online reservations are cancelled more often than others. Increased lead time of a reservation additionally amplified the cancellation probability. After all, with more days left until check-in, there naturally exists a higher probability that unexpected circumstances arise that would lead a guest to cancel their reservation.

Hypothesis 3: Percentage of the price paid in advance and the daily rate are additional determinants for the cancellation probability

We argue that the more was paid before check-in and the more is paid in general, the more reluctant will a guest be to cancel, because, as long as partial refunds are granted, more money can be refunded. However, because all travelers had their bookings made by companies, that effect, while possibly existent, might be less pronounced for Hotel California.

BACKGROUND

Our analytics and modelling pipeline inherits most of its robustness from statistical methods that ensure that findings are actionable and significant, not all of which were introduced as part of our Machine Learning 1 course. The following introduces and explaines these methods and the contexts in which we used them.

JARQUE-BERA GOODNESS-OF-FIT TEST

The Jarque-Bera test statistic indicates if any univariate distribution posesses the skewness (measure of balance in distributions) and kurtosis (measure of tailedness in distributions) that would be expected under the assumption of a normal distribution. Perfect normal distributions have a skewness of zero and a kurtosis of three. However, these may differ slightly and the distribution may still be considered gaussian. The Jarque-Bera introduces statistical significance to answer the question of precisely how much deviation from perfect gaussianity may be acceptable.

Its test statistic ranges from zero to infinity. If it becomes larger, then either skewness or kurthosis deviates from its gaussian optimum. Thus smaller p values indicate larger deviations from gaussianity.

A plethora of methods, many used in our pipeline, assume gaussianity of variables. These include Pearson correlations, two-sample t tests, Welsh tests and z-scores, which are used in standard scalers. Hence, testing if the gaussianity assumption holds for our variables was essential for the applicability of the methods that eventually drive and enable insight.

TWO-SAMPLE PERMUTATION TESTS

Measuring the level of association between a binary and a continuous variable is often done by computing a mean or median from the continuous variable for each value of the binary variable and then looking at the difference in means or medians across these two groups. To answer the question of how likely it was for that group mean/median difference to have arisen out of pure chance, statistical tests are needed. Parametric statistical tests, i.e. those that assume a distribution of the test statistic under the null hypothesis, often rely on the assumption of gaussianity of the continuous variable. Non-parametric tests, like the monte carlo permutation test, circumvent that assumption by randomly generating a distribution of the test statistic under the null hypothesis.

The two-sample permutation test randomly redraws from the binary variable with replacement of values (bootstrapping) a prespecified number of times and records the subsequent group mean/median difference every time. Thus, the implicit null hypothesis is that of independence of the two variables. After all, when every observation of the continuous variable has one of the two possible values of the binary variable assigned to it randomly, then no association between the two

can exist. The test then calculates p values by comparing the actually observed group mean/median difference with those generated by randomly bootstrapping the binary variable. Thus, smaller p values indicate higher chances of association between the two variables.

SCALED MEAN DEVIATION

Scaled mean deviation is a generalization of the idea of group mean differences. It computes the average deviation from the overall mean of a continuous variable when calculating means for each group of a (binary or multiclass) categorical variable. It then normalizes these mean deviations using the continuous variable's overall mean.

For continuous variable x and another categorical variable's unique level i with n unique levels in total, we want to compute the following:

$$SMD = \frac{1}{n} \cdot \sum_{i \in [1,n]} \frac{|\bar{x_i} - \bar{x}|}{\bar{x}}$$

We used this metric to create comparability for the associations between categorical and continuous features.

FISHERS EXACT TEST

The fishers exact test statistic indicates the confidence that is to be had in the association between two binary nominal variables. It does so by assuming that the margins of the contingency table generated by the two variables are fixed, i.e. the proportions of observations in each category remain the same for each variable when reshuffling the vectors.

The test assumes the null hypothesis that the true odds ratio is equal to one. The odds ratio is the ratio of proportions of values in the contingency table. When it is equal to one, then the variables do not associate.

We used the test to validate hypotheses about the association between binary features and the response variable "Cancelled".

CRAMERS V

Cramers V is an association measure for two multiclass nominal variables. We used it, because tests like the Chi-squared independence test only give a binary response of either rejecting or failing to reject the null hypothesis of independence. Using Cramers V allowed us to make more differentiated statements about which features are redundant during our feature selection process.

METHODOLOGY

The process from loading the data to building deployable models that we devised for our pipeline consists of three integral steps: data preprocessing, feature selection and modelling. Note that, in this section, our exploratory analysis is considered as extraneous to the modelling work.

PREPROCESSING

Predictive models, especially non-parametric ones, rely heavily on the quality of the data that is provided. As models get more abstract and powerful, the influence that the modeler can wield decreases. Thus, so does his ability to make up for poorely prepared data.

Properly preparing the data includes a few crutial steps. Not all of them were demanded by our data. For instance, our data set had no missing values and was already numerically encoded, limiting the necessitated work to the following steps.

Enforcement of integrity constraints

Datasets are pulled from transactional or analytical databases, whose quality may not be assumed without verification. Thus, we explored the existence of non-sensical data points that should not have been in these databases to begin with. We confirmed the absence of duplicated rows in the data and performed a variety of checks that aimed at verifying that the values for each variable were consistent with what the variable was meant to measure (e.g. identifying negative values for variables that measure age, number of guests, income, etc.).

Splitting the data

There are multiple methods applicable for splitting the data. In our current course, the Hold-Out Method, k-Fold Cross Validation and Leave-One-Out Cross Validation were covered.

The need to cross-validate the models we train on the training set(s) decreases with increasing data volume. After all, we expect the model to be more consistent when it was trained and tested on larger amounts of data. Our data set includes almost 14,000 observations. With that volume of data, we are confident that using the Hold-Out Method produced adequate results and successfully limits the run-time of our pipeline.

Feature scaling

Models that rely on distance measures and norms produce inaccurate predictions when variables have dissimilar value ranges, giving rise to the necessity to converge to similar upper and lower bounds of value ranges across features. There exist a variety of methods for scaling data. We will consider three of them, the Standard Scaler, the Min-Max Scaler and the Robust Scaler.

The Standard Scaler requires features to follow gaussianity. To validate if we could reasonably have made that assumption, we performed a statistical goodness-of-fit test for gaussianity - the Jarque-Bera test. Its p values lead us to reject the null hypothesis of gaussianity for all features, dissipating our confidence in the accuracy of using the Standard Scaler.

The Min-Max Scaler does not rely on gaussianity, similar to the Robust Scaler. However, because it divides every value by the existing value range of the respective feature, it does not mitigate the effects of outliers. The Robust Scaler does that better, as it divides values by the interquartile range.

The choice between the remaining two scalers thus initializes a discussion about the presence and importance of outliers. For features with a low number of unique values, outlier presence can simply be detected by looking at value counts for each feature. For quasi-continuous or continuous features

that have many unique values, it makes sense to look at values that possibly lie either above the upper quartile or below the lower quartile, where we control the definition of what an outlier is with some threshold t. The upper bound ub and lower bound lb for what constitutes an outlier will hence be defined as follows.

$$ub = q_{0.75} + t \cdot (q_{0.75} - q_{0.25})$$
$$lb = q_{0.25} - t \cdot (q_{0.75} - q_{0.25})$$

The default choice of t is usually 1.5, as used in boxplots. We sequentially increased t to get an idea about the extremity of outliers in continuous variables.

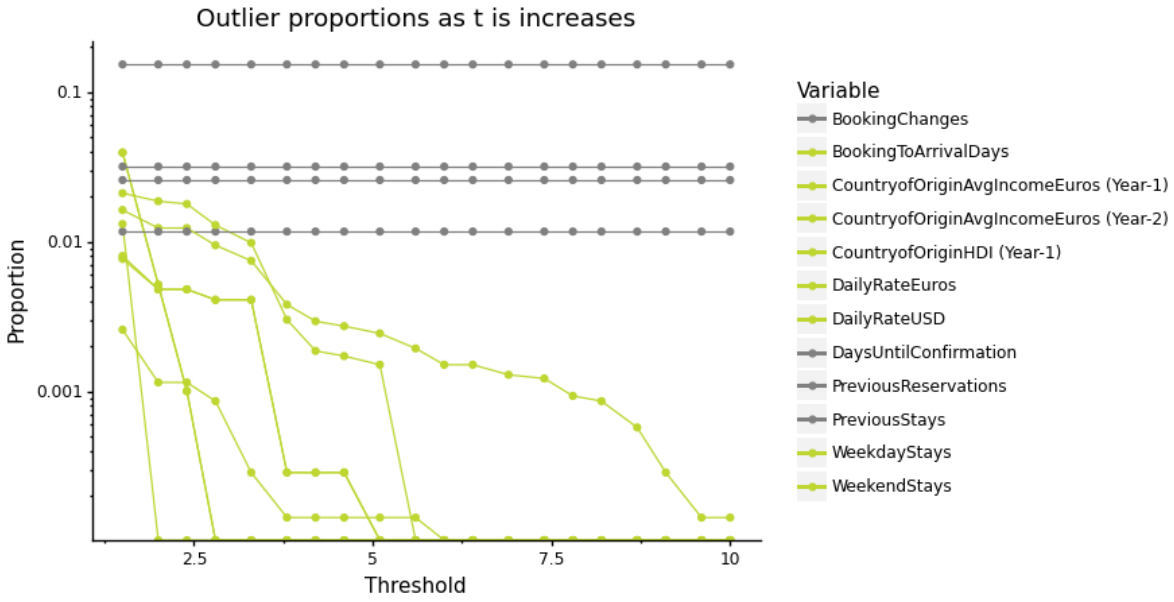

Figure 1 – Illustrative figure

We observe that even variables with more than 10 unique values have outlier proportions of above 10% in one case. “BookingChanges“ has an outlier proportion of 15%. In general, four continuously-treated features have stable outlier proportions beyond t = 10 . That was caused by highly left-sided distributions and thus low upper quartiles, as the table below indicates.

Table 1 – Illustrative table

Variable	Lower Quartile	Upper Quartile
DaysUntilConfirmation	0.0	0.0
PreviousStays	0.0	0.0
PreviousReservations	0.0	0.0
BookingChanges	0.0	0.0

Since the interquartile range is zero in all four cases, outlier proportions would never decrease, for any choice of t.

For continuous features, we conclude that four of them have outliers that - while existent - are also tough to measure using our devised methodology. There are also other features that have outliers which are up to ten interquartile ranges more extreme than their respective upper/lower quartile. These observed patters, in our opinion, yield the necessity to treat the outliers before modeling, leading us to use the Robust Scaler. It is noteworthy that the Robust Scaler does not equalize value ranges across features like the Min-Max Scaler does, instigating arguments about its effectiveness for models reliant on distance measures. Another option we considered was to binarize left-heavily distributed features to alleviate continuous outliers and then apply the Min-Max Scaler. However, the information and variance loss incurred through binarization yields worse model performance than models trained on data to which the Robust Scaler was applied.

Feature engineering

In order to find patterns in data that might be easily overlooked, it is helpful to create new variables based on other available features. That way, it may become easier to access patterns in the data that should make sense in qualitative a-priori considerations. Below, we defined a few new variables that we believe could be helpful in predicting cancellations.

- Total nights spent for a reservation are the addition of weekend nights and weekday nights: *TotalNights = WeekendStays + WeekdayStays*
- The number of underaged people on each reservation are the addition of children and babies on that reservation: *NKids = Children + Babies*
- The number of people on each reservation are the addition of underaged people and adults on that reservation: *NPeople = NKids + Adults*
- Customers are affiliated on their first reservation if the reservation is both their first reservation and it was made using their affiliate login: *AffiliatedOnFirst = (FirstTimeGuest == 1 && AffiliatedCustomer == 1)*
- Reservations experienced floor changes if the floor they were assigned differs from the floor they requested a room on: *FloorChange = 1/True, if (FloorAssigned - FloorReserved != 0)*
- The OriginCountryID is a unique identier for each origin country without providing the country name

Note that we refrained from creating features that are simply binned or binarized versions of one other individual feature. We did so because such new features would neither make our analysis easier, nor would they add new information to the data set, all while losing variance of the initial feature. A loss of variance translates into a loss of information.

FEATURE SELECTION

More data given to a model for training does not categorically translate into better model performance. To keep run time of the employed algorithms manageable and to optimize model performances, we worked to strategically subset our available features that are passed on to the modelling stage. We split the subsetting of features into two phases, one where features were continuously eliminated after respective tests have been performed (“gradual subsetting phase”) and one that encompassed a wide array of tests and selection methods that are all voting to eliminate a

set of features after their collective application ("voting phase"). Within these two phases, we used the following methods:

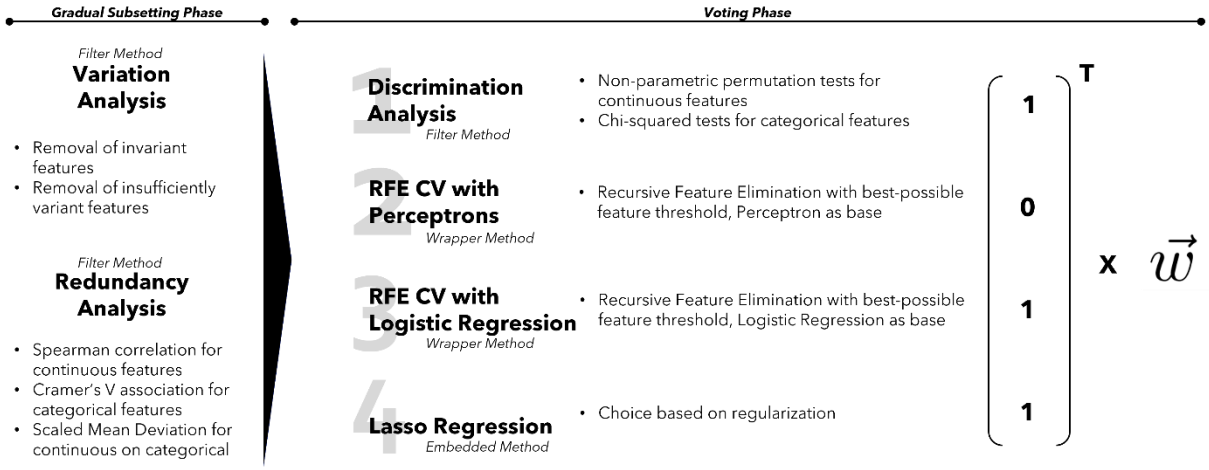

Figure 2 – Illustrative figure

Process of subsetting the features used for modeling

Gradual subsetting phase

Since variation translates to information, invariant features do not contribute to model performance and may be eliminated without further considerations.

Highly correlated features may dampen model performance. It is thus crutial to identify sets of redundant features and keep only those that have the highest discriminant potential on the response variable. Since different correlations measures suit features with different characteristics, we performed the redundancy analysis in three stages.

First, we measured correlations between pairs of continuous features, using Spearman's correlation coefficient due to our features lacking gaussianity. After sets of redundant features were identified, we used scaled mean deviation (SMD) for each continuous feature in the redundancy set to find the feature with the highest discriminant power on the binary response variable.

In the second stage, we measured associations between pairs of categorical features, using Cramer's V association coefficient. After redundancy sets have been identified, discriminant power on the response variable was measured also using Cramer's V.

In the third stage, we measured associations between continuous and categorical variables. The association was measured using SMD measures. The methods used to determine discriminant power on the response variable were either Cramer's V or SMD, depending on the variable at hand.

Voting phase

The discrimination analysis in the voting phase was conducted analogously to the estimation of discriminant power during the redundancy analysis.

As a second and third step, we used a method called Recursive Feature Elimination (RFE). RFE starts by building a model with all available features and removes the least important feature from the model sequentially, until a prespecified number of features is reached. The remaining features should be kept, its set difference to the set of all features should be discarded. To find the optimal number of features to keep, we recorded the performance of the optimal model depending on the number of features to keep for two different base models.

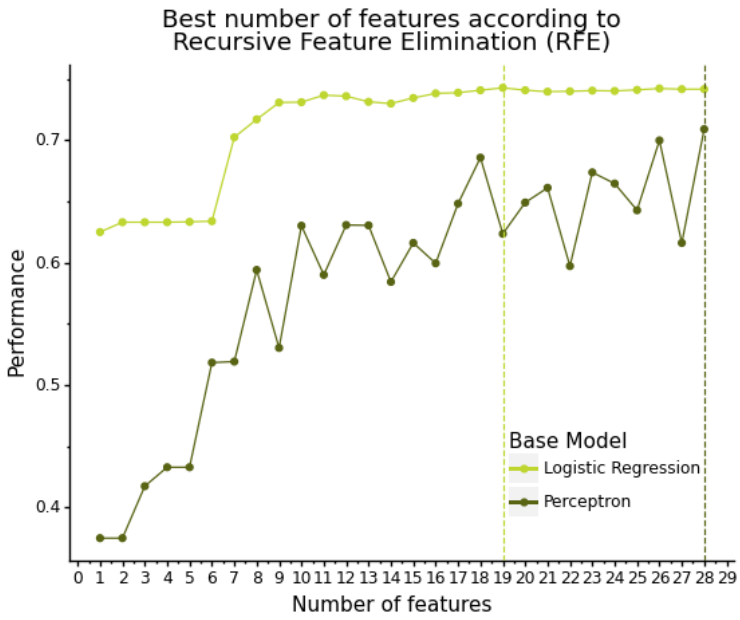

Figure 3 – Illustrative figure

We observed that the Perceptron RFE performs best when all features are kept. It is thus not as relevant when it comes to selecting a subset of features. Hence, we give its vote less weight.

The fourth method we used in the voting phase is Lasso Regression, which is able to assign importance measures to features while adding a regularization penalty to unimportant features. Features with importance of zero should then be discarded.

We let the Perceptron's voting power be compensated by giving more weight to the discrimination analysis, because it is a straight-forward approach that distinguishes based on easily understood methodologies. After weight-averaging the decisions and cutting off at 0.5, we eliminated four additional features.

MODELING

We hypothezise that ensemble learners and Neural Networks will perform better on our data than singular linear, instance-based or tree-based classifiers, as they tend to perform better at filtering out noise in data. To validate our hypothesis, we built one Stacking Classifier on singular models (Decision Tree, Logistic Regression and kNN Classifier) and one for ensemble models and Neural Networks (Boosting Classifier, Bagging Classifier, Random Forest, Multi-Layer Perceptron). We optimized the hyperparameters for each model before training the respective Stacking Classifier.

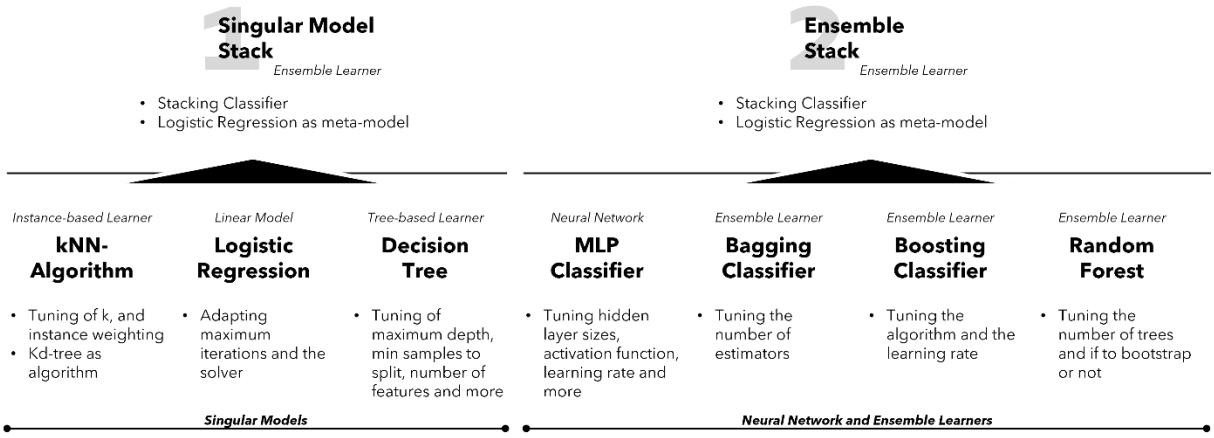

Figure 4 – Illustrative figure

Overview over modeling architecture for Hotel California

In total, we had seven base models to optimize as inputs for two Stacking Classifiers. For Decision Trees and the MLP Classifier, we optimized hyperparameters on a trial and error basis without visualization, since too many parameters had to be adjusted for visual pattern detection to be possible.

kNN-Algorithm

As an algorithm, we chose to use a kd-tree, since kd-trees partition the feature space before distances are computed within it, thus limiting the combinatorial surfeit of distances the algorithm has to compute.

When making predictions, the weighting of an instances neighbors influences the likelihood of correct predictions. We assumed distance-based weighting to be more reasonable, but tested models both for distance-based and uniform weights.

The choice of k, i.e. the number of neighbors for the algorithm to consider when making predictions, cannot be optimized deterministally a-priori. Thus, we built multiple models and varied k each time.

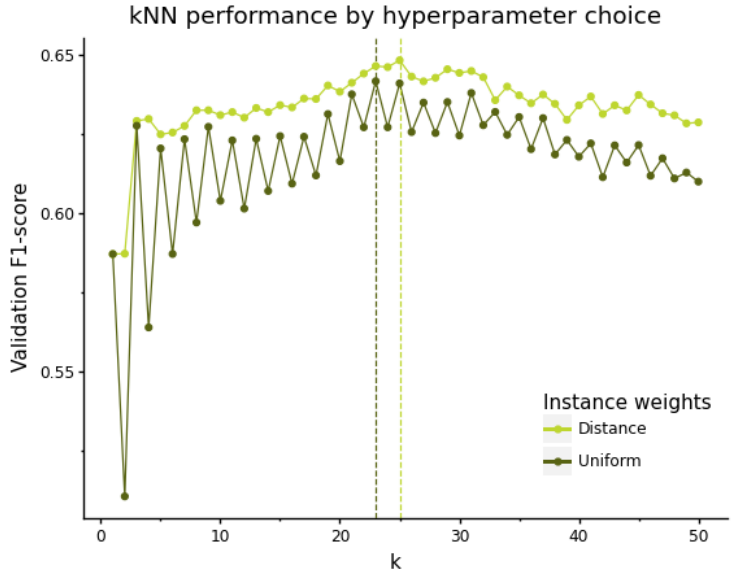

Figure 5 – Illustrative figure

We observed that distance-weighting indeed produces better models. Then, the optimal choice for k is 25. Thus, our final model for the subsequent stack considered 25 distance-weighted neighbors and used a kd-tree to partition the feature space.

Logistic Regression

In total, we trained a 100 Logistic Regressors, having varied the maximum number of iterations 20 times as well as the solver that is used. The optimal model allowed for a maximum of 240 iterations and used the solver "lbfgs".

Decision Tree

For decision trees, we tuned the tree's maximum depth, the number of features it may use, the minimum impurity decrease, the minimum fraction of instances in each leaf as well as the splitting criterion (Gini impurity or entropy). We did not consider all combinations for run time purposes, but rather tried parallel sets of possibilities, the optimum of which we kept as the best model for the single model stack. The best model had a maximum depth of ten levels, used at most 70% of available features, split according to Gini impurities, could have any fraction of instances in each leaf and did not need any impurity decrease to split further.

MLP Classifier

We additionally added a feed-forward Multi-Layer Perceptron to the Ensemble Stack. The hyperparameters we optimized were the alpha of the "Adam" solver, hidden layer sizes, activation function, maximum number of iterations as well as the learning rate initialization. The optimal "Adam" MLP Classifier had an alpha of 0.02, two hidden layers, the first with ten neurons, the second with nine neurons, a sigmoid as its activation function, an initial learning rate of 0.01 and a maximum of 1000 iterations.

Bagging, Boosting and Random Forest Classifiers

All three methods are ensemble learners and have intersecting but not equal sets of hyperparamers. Thus, we optimized each model separately. For the Boosting Classifier, we tuned the learning rate, and the algorithm. Both for the Bagging Classifier and the Random Forest, we tuned the number of estimators. For Random Forests, we additionally built different models for bootstrapped samples and non-bootstrapped samples.

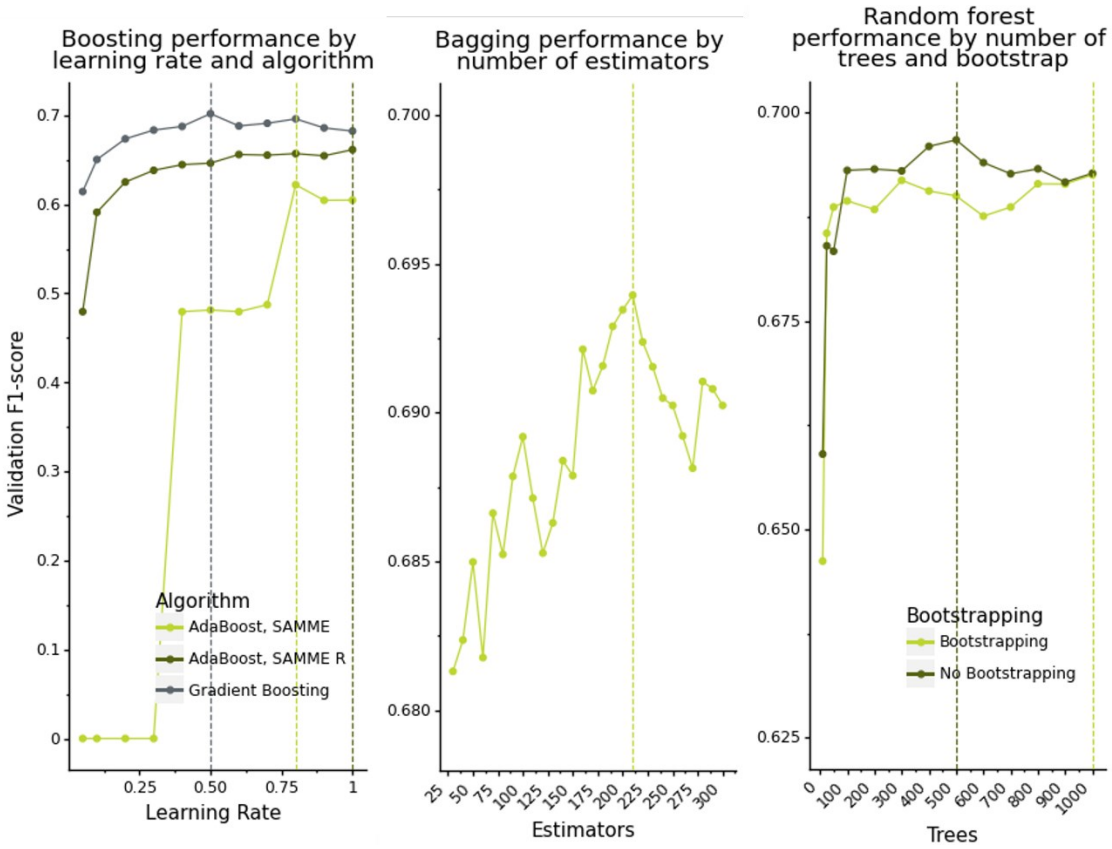

Figure 6 – Illustrative figure

The plot indicates that Boosting Classifiers perform best on Gradient Boosting and with a learning rate of 0.5, Bagging Classifiers perform best with 210 estimators and the Random Forest performs best when samples are bootstrapped and 500 trees are considered.

RESULTS

HYPOTHESIS 1

Before exploring major determinants for cancellation probability, we had to confirm or dismiss the implicit assumption that booking and cancellation behavior is not cyclical. If they were to be cyclical, then further analyses would have to be stratified by the dimension of time.

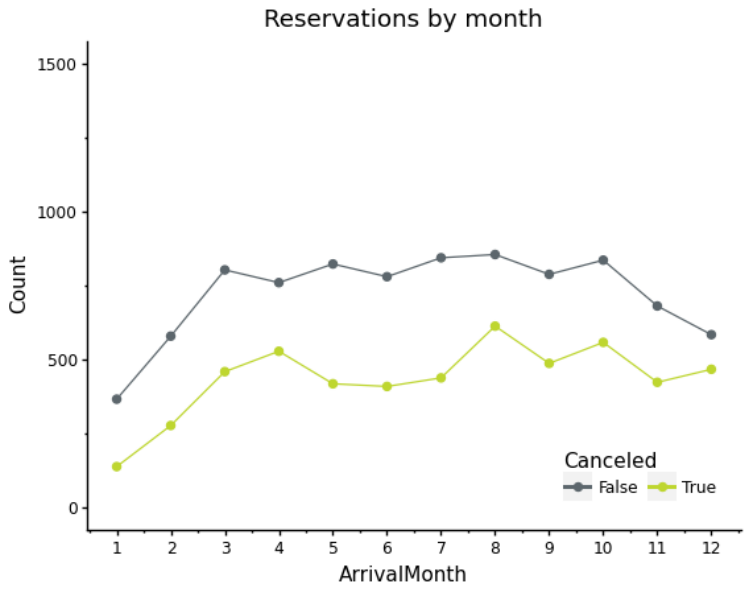

Figure 7 – Illustrative figure

Figure 7 confirms hypothesis 1. While bookings overall decrease in the beginning and end of the year, they remain stable throughout the remainder of the year. The biggest cancellation percentage can be observed in May.

HYPOTHESIS 2

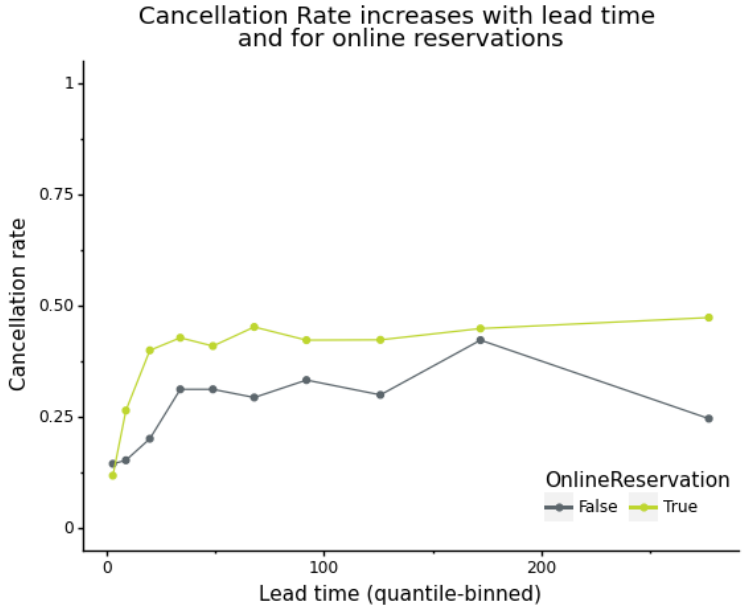

Figure 8 – Illustrative figure

Figure 8 displays the cancellation rate for Hotel California as the lead time increases, differentiated by the booking channel (online or not). Note that the lead time, measured in days, was binned before plotting. Binning was mainly done for the purpose of better visual comprehension. The binning was based on quantiles, i.e. all bins are populated similarly, because lead time has a left-heavy distribution, yielding bins for high lead times to be so scarcely populated for offline reservations

otherwise, such that differences in cancellation rates would have had no chance to be statistically significant without quantile binning.

Interestingly, the plot not just confirms our primal expectations, it additionally reveals that, for short lead times, the mode of reservations had virtually no impact on cancellation decisions, i.e. if a guest books very spontaneously, it does not matter how they booked. Additionally, we observe decreasing marginal cancellation rates for increasing lead times and a degree of concavity in the plotted line. Cancellation rates increase the quickest up to around 50 days of lead time. After that, additional days of lead time have small to negligible effects on the cancellation rate.

The p value of the association between lead time and the cancellation rate is 0.002, with a group mean difference in lead time of 18.8 days and a median difference of 22 days of lead time for cancelled reservations versus uncancelled reservations. The p value for the association between the mode of reservation and the cancellation rate is $8.022 \cdot 10^{-50}$.

The findings of Falk and Vieru are prevalent and statistically significant in our data. We additionally hypothesized that the amount a guest paid influenced their reluctance to cancel a reservation (hypothesis 3).

HYPOTHESIS 3

When stratifying by cancelled reservations, we observed that the percentage that was paid in advance is higher when the booking was cancelled, supporting our hypothesis 3. A similar pattern is detectable for the overall daily rate that guests were obliged to pay.

Table 1 – Illustrative table

Cancelled	Percentage paid in advance	Daily rate in EUR
0	0.0015	105.12
1	0.0273	113.03

Both associations are statistically significant at a permutation test p value of 0.00199 each.

MODEL PERFORMANCE

Our best performing model is a Stacking Classifier based on Bagging, Boosting and Random Forest Classifiers. On unseen test data, it has an F1-score of 0.79234. We have trained and optimized 10 models in total.

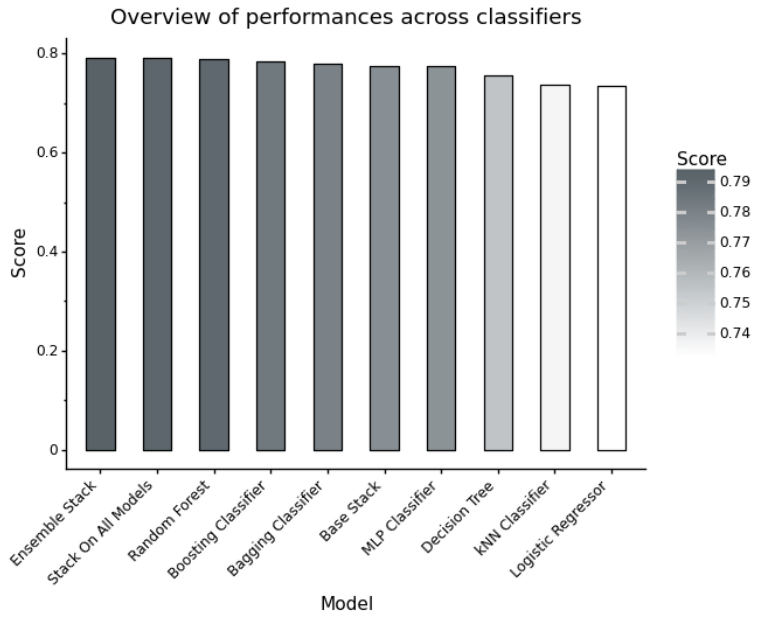

Figure 8 – Illustrative figure

We additionally observe that singular base models perform the worst, as was expected during our modelling work. Ensemble Classifiers perform the best, with Stacks on Ensemble Classifers outperforming all models. Thus, we recommend the Ensemble Stack for operational deployment.

DISCUSSION

Our research has shown that online reservations with long lead times and high proportions of high prices being paid in advance are most prone to cancellations. These findings have wide ranging implications on Hotel California's pricing model, refund policy and overbooking strategy. They would suggest incentivizing spontaneous reservations, a distinction of refund policies by mode of reservation and a facilitation of more flexible payment plans.
In addition, our findings are congruent with third-party research, specifically with the conclusions of Falk and Vieru, who, among other factors, identified lead time and mode of reservation, as prime discriminants for cancellations.

Furthermore, the design of refund policies will influence future optimizations of predictive models. The refund policy dictates how costly it is to falsely predict cancellations. If a guest is predicted to cancel their reservation but actually does not and another guest already booked the room, then Hotel California would have to cancel one of the two reservations. If a guest cancels their reservation without the cancellation being predicted by the model, then revenue is lost on short notice. The terms of Hotel California's refund policy thus dictate how costly false positives are compared to false negatives, which influences the metric that is to be used when evaluating model performances.
The Ensemble Stack that we recommend for deployment has higher precision than recall, i.e. it performs better on false positives than false negatives, meaning that it predicts cancellations that are none less often than it predicts an actual check-in that ends up being cancelled.

CONCLUSION

Our introductory remarks presented the importance of an overbooking strategy as a vital contributor to profit-optimality for Hotel California. In their attempt to mitigate profit elasticity for changes in cancellation rates, hotel management should use the Stacking Classifier on ensemble models that we trained for day-to-day cancellation predictions, while leveraging our qualitative research findings to proactively reduce cancellation rates in the future. That proposed dual-use of our findings will enable Hotel California to both formulate a comprehensive overbooking strategy and minimize the importance of such a strategy to begin with.

Future work will have to include iterative optimization of the models we have built. For that, close collaboration with Hotel California will be essential, especially in developing refund policies and overbooking strategies that help to associate true costs to false positives and false negatives. Through these initiatives, we believe that Hotel California will be able to dampen cash flow volatility arising from cancellations and boost profits in the long run.

REFERENCES

[1] Martin Falk, Markku Vierru. Modelling the Cancellation Behaviour of Hotel Guests. 2018

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa