Hotel Reservation Cancellation Prediction

Final Report

Ruochen Wang

Executive summary

In the ever-evolving landscape of the hospitality industry, predicting hotel reservation cancellations has emerged as a crucial challenge for hoteliers. With a constant influx of bookings and the financial implications of cancellations, it has become imperative to leverage advanced data analysis techniques to address this issue effectively. This project delves into the realm of data-driven insights by employing Random Forest for feature selection and Neural Network for prediction. By harnessing the power of these algorithms, we aimed to provide hotels with a cutting-edge solution that enables proactive decision-making, optimizes revenue, and enhances customer satisfaction.

Introduction

The goal of my project is to predict hotel Hotel Reservation Cancellation. Reservation cancellation is a common issue that businesses in the hospitality industry face. It not only affects the profitability of hotels, but it also leads to inefficiencies in resource allocation, as rooms that were originally reserved end up being unoccupied. Therefore, it is crucial for hotels to have an accurate prediction of reservation cancellations to adjust their inventory and pricing strategy accordingly.

The project is not only useful for businesses in the hospitality industry, but it also has broader implications for other industries that face similar prediction problems, such as airline ticket cancellations, online shopping cart abandonment, and appointment no-shows.

The aim of this project is to analyze reservation cancellation data and develop a machine learning model to predict cancellation. The analysis and model will help hotel managers optimize their revenue and resource allocation strategies.

Related Work

The topic of hotel cancellation prediction has many related studies, including academic papers and some projects. As I choose a Kaggle dataset, some Kaggle users have also published related analyses. I will present both references in academic papers and in relevant projects. Expanding what I included in the proposal, I'll add one academic paper and one relevant project here

1. Academic Study

Performance analysis of machine learning techniques to predict hotel booking cancellations in hospitality industry by Md. Shahriare Satu; Khair Ahammed; Mohammad Zoynul Abedin

In the research, this data was generated by Antonio, Almeida, and Nunes that contained two hotel booking information like resort(H1) and city hotel (H2). It was gathered from their databasesby executing Transact-Structured Query Language (TSQL)query. All instances pertaining hotel or customer identificationnumber were removed. Then, we combined these datasetsalong with 31 variables where 40,060 observations of H1and 79,330 observations of H2 were found. However, a cityhotel (41.90%) had shown higher cancellation rate than resorthotel (27.69%).

The author applies data preprocessing techniques to clean and manipulate the input dataset, including handling missing values and removing irrelevant columns. They also perform feature transformation by scaling, standardizing, and normalizing the features, generating multiple transformed datasets. Additionally, feature selection methods such as Correlation-Based Feature Selection (CFS), Info Gain Attribute Evaluation (IGAE), and Gain Ratio Attribute Evaluation (GRAE) are used to identify relevant features and create subsets of data. These steps aim to improve the generalization and performance of the machine learning model.

The author employs various classification algorithms, including Gradient Boosting (GB), Random Forest (RF), Xgboost (XGB), Decision Tree (DT), Logistic Regression (LR), K-nearest Neighbors (KNN), and Gaussian Naive Bayes (GNB), to predict booking cancellations based on individual customer records. These classifiers are widely used in different machine learning projects and datasets. The performance of each classifier is analyzed to determine their effectiveness in building an efficient model. The evaluation is done using 10-fold cross-validation to ensure unbiased results and mitigate overfitting issues.

The author uses various evaluation metrics to assess the performance of the machine learning models and determine the best one. The metrics employed include Accuracy, AUROC (Area Under the Receiver Operating Characteristic curve), and F-Measure. Accuracy measures the correctness of the classifier by considering true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). AUROC quantifies how well the classifiers distinguish negative classes and is calculated using true positive rate (TPR) and true negative rate (TNR). F-Measure combines precision and recall, where precision represents the number of similar instances divided by the total number of records, recall represents the number of correctly classified similar records divided by the total number of instances, and F-Measure calculates the harmonic mean of precision and recall. These metrics are utilized to evaluate and compare the performance of the models.

The results of the research showed that different machine learning algorithms performed well for predicting hotel booking cancellations in different subsets of the dataset. Gradient

Boosting (GB) and Logistic Regression (LR) were found to be the best classifiers for the primary dataset and feature-transformed datasets. Xgboost (XGB) also achieved high accuracy, AUROC, and f-measure in the feature-transformed datasets. Feature selection techniques were effective in improving the performance of the models by reducing irrelevant features. However, feature transformation techniques did not provide significant improvements in the results. The study suggests that further research can be conducted using advanced machine learning and deep learning approaches, as well as gathering more data from diverse sources, to build more accurate predictive models for hotel booking cancellation.

This project offers several important lessons. It highlights the significance of data preprocessing, the selection of classification algorithms, and the use of evaluation metrics for assessing model performance. The project also emphasizes the importance of feature selection techniques in improving model efficiency. It provides insights into potential future research directions and opportunities, such as the use of advanced machine learning and deep learning approaches. Overall, the project serves as a valuable resource for learning about and applying machine learning techniques in predictive analytics.

2. Kaggle project

Hotel Booking Demand (EDA+Cancelation prediction) by Kaggle user ARMAN AKBARI
https://www.kaggle.com/code/armanakbari/hotel-booking-demand-eda-cancelation-prediction

The dataset is “Hotel booking demand" dataset by Kaggle, which is exactly what I am going to use in my project. The author performs several data preprocessing steps, including handling missing values, encoding categorical variables, and dropping irrelevant columns. The project provides an extensive exploratory data analysis, examining various aspects of the dataset. It explores the distribution of bookings, cancellation rates, booking patterns over time, and correlation between different features. machine learning models applied are logistic regression, k-nearest neighbors (KNN), decision tree, random forest, and gradient boosting classifiers.

The project evaluates the performance of the models using metrics such as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC).

For results of the project, the author presents the results of model predictions, including the accuracy, precision, recall, F1 score, and AUC-ROC values. Additionally, he discusses the importance of different features in predicting reservation cancellations.

This research provides me with some insights into the data preprocessing for my project. I learned how to deal with missing data and categorical data. In addition, I also learned performance evaluation metrics. Metrics such as accuracy, precision, recall, F1 score, and area

under the receiver operating characteristic curve (AUC-ROC) helps compare the effectiveness of different models.

Description of solution:

A. Data Source

The dataset I use for our project is the "Hotel Booking Demand" dataset in Kaggle. https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand. The data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. https://www.sciencedirect.com/science/article/pii/S2352340918315191. The dataset contains 32 variables and more than 119390 records. Each observation represents a hotel booking, including information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. The datasets comprehend bookings due to arrive between the July 1st of 2015 and the August 31st 2017, including bookings that effectively arrived and bookings that were canceled.

This dataset is not the one I mentioned in my project proposal. I planned to use "Reservation Cancellation Prediction" dataset provided on Kaggle. Although the “Reservation Cancellation Prediction” provides cleaner data and is “designed” for the reservation cancellation analysis, Hotel Booking Demand gives more comprehensive information. It has much more variables and records, which helps model training. Correspondingly, I did more data cleaning and feature filtering work due to the increased data complexity.

B. Software

I used Python and various libraries including Pandas, NumPy, Matplotlib, and Scikit-learn to perform data cleaning, preprocessing, visualizing, and modeling. NumPy is used for numerical computations and matrix operations. Pandas is used for data cleaning, data manipulation, and data exploration. Matplotlib is used to create various types of charts, which better visualizes the data and results of my model. Scikit-learn provides a wide range of machine learning algorithms, including LogisticRegression, RidgeClassifier, RandomForestClassifier, and more. I used Scikit-learn to train and evaluate models to predict whether a reservation will be canceled or not.

C. Discription of Analysis and My Effort

I implemented the Reservation Cancellation Prediction project involves several steps:

1. Data Cleaning and Preprocessing

Firstly, I dropped column with too many NULL value, repetition, or little significance. I dropped 'arrival_date_year' because year is similar among recode and the year is a large number, which can bias the model. I dropped 'arrival_date_week_number' because it can be represented by 'arrival_date_month' and 'arrival_date_day_of_month'. I drop agent and company because I the recode has too many NULL value and these 2 features are not very important in my analysis. I dropped 'reservation_status' because I can be represented by 'is_canceled', which is the feature I need to predict.

In addition, I used LabelEncoder to transform categorical columns 'hotel', 'meal', 'country', 'market_segment', 'distribution_channel', 'reserved_room_type', 'assigned_room_type', 'deposit_type', and 'customer_type' into numerical representations. The 'arrival_date_month' column is encoded using a mapping dictionary that maps month names to corresponding numerical values. The data are prepared for further analysis and modeling by encoding categorical features, as many machine learning algorithms require numerical inputs.

Finally, I splited the dataset into 80% training and 20% testing sets. I evaluated the models using K-fold cross-validation, calculate model accuracy metrics on the test set.

2. Feature Selection

For feature selection, I performed correlation analysis and assessed the feature importance using a Random Forest Classifier on the provided dataset. To begin, I computed the correlation matrix and visualized it using a heatmap. This allowed me to gain insights into the relationships among different variables. Next, I trained a Random Forest Classifier model to determine the importance of each feature. By sorting and plotting the feature importance in a bar chart, I could easily identify the most influential features in the dataset. I chose Random Forest Classifier to determine feature importance before model training because it provides a quantitative measure of each feature's contribution to the model's predictive performance. It captures non-linear relationships and interactions, helps with feature selection, and improves model interpretability. Finally, based on the feature importance scores, I selected the top 15 features for further analysis or modeling purposes. These features are expected to have a significant impact on predicting the target variable. By analyzing correlations and feature importance, I gained a better understanding of the dataset and identified the most important features for predicting the desired outcome.

3. Model Training

For the model training, I trained various models using K-Nearest Neighbors, Logistic Regression, Ridge Regression, Neural Network, and Support Vector Machine algorithms. Since I used Random Forest Classifier model to implement feature selection, I did not include it in the model training part. For each model, I used GridSearchCV to search for the best parameter value from a predefined range and evaluated the model's performance using cross-validation. By plotting the mean training and validation scores for each parameter value, I visualized the performance of the algorithm with different parameters and got a better idea

when choosing paramethers. Then, I used model with the best parameter, I made predictions on the test set and calculated the accuracy score and classification report to evaluate its performance. Additionally, I generated a confusion matrix to visualize the predicted labels against the true labels, which gave me further insights into the model's predictive capabilities, showing the distribution of predicted and true labels.

4. Data Visualization

Data visualization accompanies me every step of my project.
Firstly, I created a countplot to visualize the distribution of cancellations is useful for understanding the balance or imbalance of the target variable in my dataset. It provides insights into the proportion of canceled and not-canceled instances.
In the feature selection, I trained a Random Forest Classifier model to determine the importance of each feature and plotted the feature importances in a bar chart so that I could easily identify the most influential features in the dataset.
For the model training, I used GridSearchCV to search for the best parameter value in each model. I plotted the mean training and validation scores for each parameter value to visualize the performance of the algorithm with different parameters and got a better idea when choosing paramethers. What's more, I visualized confusion matrixes for every model I chose to get a better idea of the performance of the performance.

5. Decide the Final Model and Make Predictions on Test Data

After deciding on the final model by step 4, I'll make a prediction on the test data provided. Details of the result are in section E. Analysis of Results below.

D. Model Evaluation

For feature selection, the benchmark I use is the feature importance scores of Random Forest Classifier. Before training my model, I spited the data set into 2 parts, train and test sets. I train the model using the train set and evaluate my model using the test set. For each model, I used GridSearchCV to search for the best parameter value from a predefined range and evaluated the model's performance using cross-validation. Then, I used model with the best parameter, I made predictions on the test set and calculated the accuracy score and classification report (include precision, recall, f1-score, support) to evaluate its performance. Additionally, I generated a confusion matrix to visualize the predicted labels against the true labels.

E. Analysis of Results

Based on the feature importance scores of Random Forest Classifier, the top 15 features or modeling purposes are 'lead_time', 'deposit_type', 'country', 'adr',

'arrival_date_day_of_month', 'market_segment', 'total_of_special_requests', 'arrival_date_month', 'stays_in_week_nights', 'previous_cancellations', 'assigned_room_type', 'customer_type', 'stays_in_weekend_nights', 'required_car_parking_spaces', 'booking_changes'.

The results for models models:
K-Nearest Neighbors (KNN) with K=7: Accuracy: 0.785, Macro-average F1-score: 0.76
Logistic Regression: Accuracy: 0.770, Macro-average F1-score: 0.74
Ridge Classifier with the best alpha (1000.0): Accuracy: 0.777, Macro-average F1-score: 0.73
Neural Network with the best hidden layer sizes (10, 20) and alpha (0.001): Accuracy: 0.808, Macro-average F1-score: 0.79
Linear SVM: Accuracy: 0.788, Macro-average F1-score: 0.76
(*For detailed results and analysis of each model above, see the Colab notebook*)

Comparing the models, the Neural Network model achieved the highest accuracy of 0.808, with a macro-average F1-score of 0.79. It showed good precision, recall, and F1-score for both classes, indicating a balanced performance in predicting both positive and negative instances. The Linear SVM model also performed well with an accuracy of 0.788 and a macro-average F1-score of 0.76.

Based on these results, the Neural Network model with hidden layer sizes of (10, 20) and an alpha value of 0.001 seems to be the best-performing model for this dataset. It demonstrates higher accuracy and better overall performance compared to other models. the model correctly predicted the class labels for approximately 80.84% of the instances.

Reflection and Outlook

A. Challenges, Solutions, and Tread-offs

1. Dataset. In the project proposal, I planned to use "Reservation Cancellation Prediction" dataset provided on Kaggle. However, during the analysis process, I found that this data set failed to provide plenty of features and recodes to train a decent model even if it is “designed” for the reservation cancellation analysis. So I changed to " Hotel Booking Demand " dataset, which gives more comprehensive information. It has much more variables and records, which helps model training. Correspondingly, I did more data cleaning and feature filtering work due to the increased data complexity.
2. Feature selection. Too many features may cause poor interpretability, overfitting, and multicollinearity. I planned to use Forward Stepwise or Backward Stepwise Selection for feature selection. But this increases the complexity of the project, especially after I get the result, I have to use the new feature set to adjust the previous mode. I opted for Random Forest Classifier to determine feature importance due to its ability to capture non-linear relationships, handle high-dimensional data, and provide interpretability. Although this model requires more computational resources and has higher complexity, it offers

improved performance and better feature selection capabilities compared to simpler models.

3. Limited computational resources. I planned to train Neural Network, and Support Vector Machine with ridge and polynomial kernel using GridSearchCV to search for the best parameter value from a predefined range and evaluated the model's performance using cross-validation. But given the data volume, this process requires large computational resources and it doesn't work on Colab. To solve this problem, I shrink the parameter predefined range of parameters of Neural Network and run the model on my own computer. For Support Vector Machine, I tried linear kernel with pre-fixed C instead of ridge and polynomial kernel. If I have more time and more computational resources, I would try more complex models.

B. Way to Get Final Solution

The final solution was chosen based on a combination of factors. Firstly, the feature importance analysis from the Random Forest Classifier helped identify the most influential features for predicting the target variable. This allowed for a more focused and meaningful analysis. Secondly, I considered the performance of various models through cross-validation and evaluation metrics such as accuracy, precision, recall, and F1-score. The models were trained and tuned using GridSearchCV to find the best parameter values. Additionally, I analyzed the trade-offs between model complexity, computational requirements, interpretability, and predictive performance.

C. Takeaways From the Project

1. Exploration of advanced models: By trying out more advanced models such as Random Forest Classifier and incorporating techniques like GridSearchCV for parameter tuning, I expanded my knowledge and experience with sophisticated machine learning algorithms. I gained insights into the strengths and limitations of these models and learned how to select the best model based on performance, interpretability, and computational requirements. This exposure to advanced models broadened my skill set and allowed me to explore more complex datasets and prediction tasks.
2. Model Evaluation: Throughout my project, I utilized various evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve to assess the performance of my models. I also employed techniques like K-fold cross-validation to estimate the models' generalization ability and mitigate overfitting. By analyzing these metrics and understanding their implications, I gained a deeper understanding of the strengths and weaknesses of my models and their suitability for my prediction task. This knowledge helped me make informed decisions and select the most effective models for my project.

D. Ideal data

The "Hotel Booking Demand" dataset from Kaggle provides valuable information for analyzing and predicting hotel reservation cancellations. However, there are some additional data points that would have been beneficial for a more comprehensive analysis.

1. Comprehensive Time Range: The dataset would cover a wide range of time, spanning several years or even decades. This would enable the analysis of long-term trends, seasonal variations, and the impact of external factors on reservation cancellations.
2. Diverse Geographical Coverage: The dataset would include data from various regions, cities, and countries worldwide. This geographic diversity would allow for the exploration of regional differences, cultural influences, and the impact of local events or attractions on reservation cancellations.
3. Granular Demographic Information: Detailed demographic data about the guests, such as age, gender, nationality, and socioeconomic status, would be included. This would facilitate the analysis of how different demographic groups contribute to reservation cancellations and help identify specific target audiences.
4. Fully diverse features. Some additional data points that would have been useful but are not available. These include customer reviews and ratings, travel purpose, booking source behavior, room preferences, competitive pricing, travel insurance, external events, guest loyalty program, socioeconomic factors, and sentiment analysis.

E. Further work

If I had more time to continue the project for the next 1-3 months, there are several steps I would take to further enhance the analysis and model performance:

1. Model exploration. Given more time, I would focus on exploring advanced models such as Recurrent Neural Networks, Transformers, Gaussian Processes, Deep Reinforcement Learning, and Gradient Boosting with Bayesian Optimization. These models offer specialized capabilities for handling sequential data, capturing complex patterns, quantifying uncertainty, and optimizing reward-based objectives.
2. Ensemble modeling: Building an ensemble of multiple models can often lead to improved performance. I would consider combining the predictions of different models using techniques such as majority voting, stacking, or boosting. This would help leverage the strengths of different algorithms and enhance the overall predictive power.
3. Feature engineering: I would explore the possibility of creating new features or transforming existing ones to capture additional information that might be relevant for the prediction task. This could involve deriving temporal features from dates, creating interaction terms between variables, or applying mathematical transformations to improve linearity.
4. Deployment and real-world testing: If feasible, I would deploy the final model into a real-world setting for further testing and validation. This would involve monitoring its performance over time, collecting feedback from users or stakeholders, and making necessary adjustments to ensure its effectiveness and reliability in practical scenarios.

Reference

https://www.kaggle.com/datasets/gauravduttakiit/reservation-cancellation-prediction
https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand
https://www.sciencedirect.com/science/article/pii/S2352340918315191
https://www.kaggle.com/code/alpayabbaszade/combined-feature-selection-analysis
https://www.kaggle.com/code/armanakbari/hotel-booking-demand-eda-cancelation-prediction
https://ia-institute.com/wp-content/uploads/2021/07/IAI-Journal_2.2021.pdf#page=4
Novakovic J, Turina S. Hotel reservation cancellations: analysis and prediction using machine learning algorithms[J]. ACADEMIC JOURNAL, 4.
Satu M S, Ahammed K, Abedin M Z. Performance analysis of machine learning techniques to predict hotel booking cancellations in hospitality industry[C]//2020 23rd International Conference on Computer and Information Technology (ICCIT). IEEE, 2020: 1-6.