Hotel Chain C

Business Case 2 - Predict Hotel Booking Cancellations

Andreia (20210604), Pauline (20211019), João (20211014), Tiago (20210766)

Search

TODAY'S AGENDA

Following CRISP-DM Method...

01 Introduction

02 Business and Data Understanding

03 Data Preparation and Tansformation

04 Predictive Model and Result's Evaluation

05 Business Applications

06 Conclusion and Limitation

INTRODUCTION

Business and Project overview

I. BUSINESS AND DATA UNDERSTANDING

BUSINESS CONTEXT AND OBJECTIVES

Consultant dataset

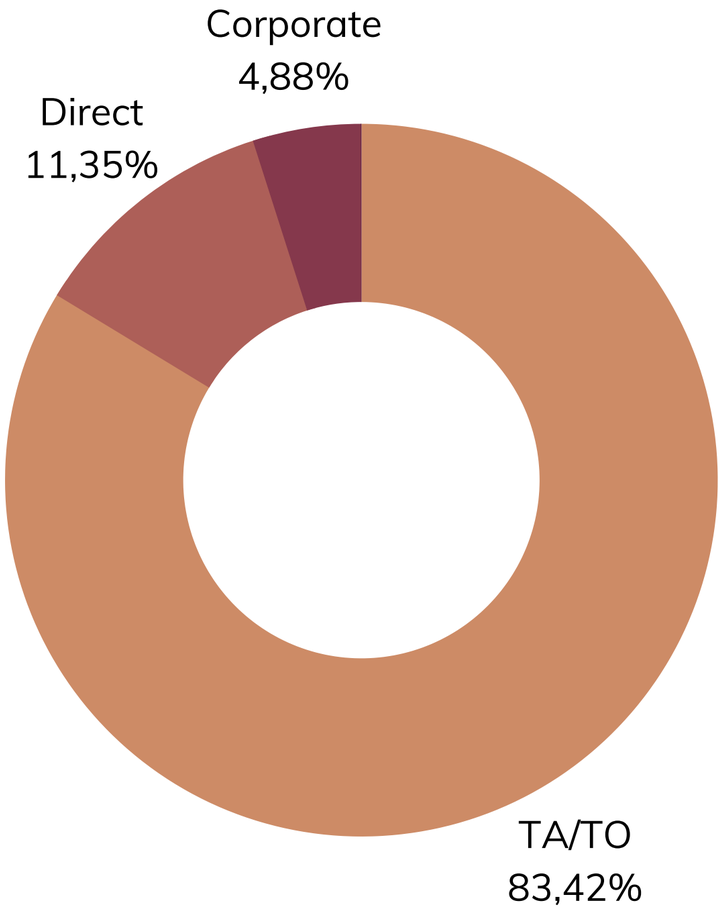
- 79330 observations;
- 31 variables;
- July 1, 2015 - August 31, 2017

Goals

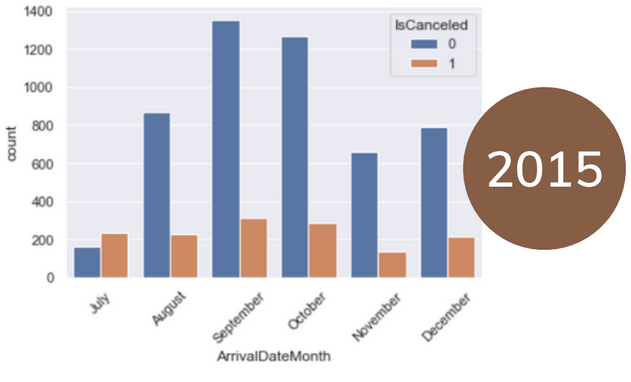
- Reduce from 42% to 20%
the cancelation rate;
- Develop a predictive model to
support the forecast of the bookings.

KEY OBSERVATIONS

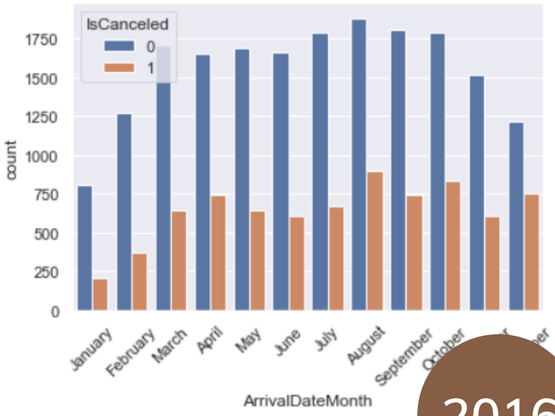
1. Distribution Channels



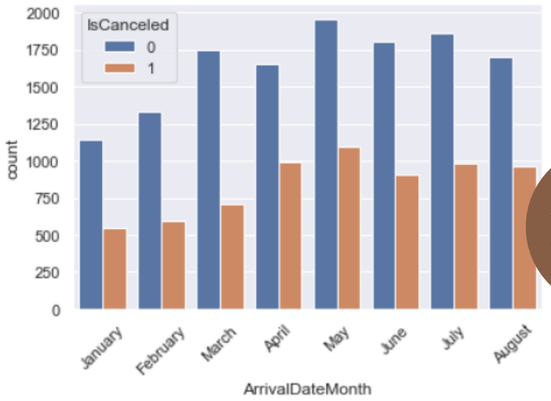
2. Cancellation throughout Years



2015



2016



2017

2. DATA PREPARATION AND TRANSFORMATION

DATA DESCRIPTION AND ASSESSING DATA QUALITY

Check for duplicates

Our Final Decision: Remove duplicates (32,6%) to prevent a biased model

Check for Missing and NULL Values

For 'Country': very small quantity so the missing values were deleted

For 'Children': missing values were replaced by 0

Data Preparation and Transformation

☑ Check for Outliers

Density-based clustering algorithm was used to detect outliers, 0,5% of the dataset was removed

☑ Feature Engineering

2 new features:

- TotalNights = StaysInWeekendNights + StaysInWeekNights
- TotalUnder18 = Children + Babies

☑ Feature Selection

The most important features to predict cancelation :

- LeadTime
- PreviousCancellations
- PreviousBookingsNotCanceled
- TotalOfSpecialRequests
- MarketSegment
- IsRepeatedGuest

3. Predictive Model and Results' Evaluation

1. Implementing Predictive Models
2. Performance Criteria 1: Cross Validation results
3. Performance Criteria 2: False Positives VS False Negatives

1 2

Implementing Predictive Models

First consideration

Should we consider the Time and use TimeSeries Split ?

Which algorithm for a binary Outcome

The 3 bests: Random Forest, Catboost and XGBoost.

Split for imbalanced Dataset

Stratified K-Fold

Find optimal parameters

Halving Grid Search

1 2

Cross Validation Results

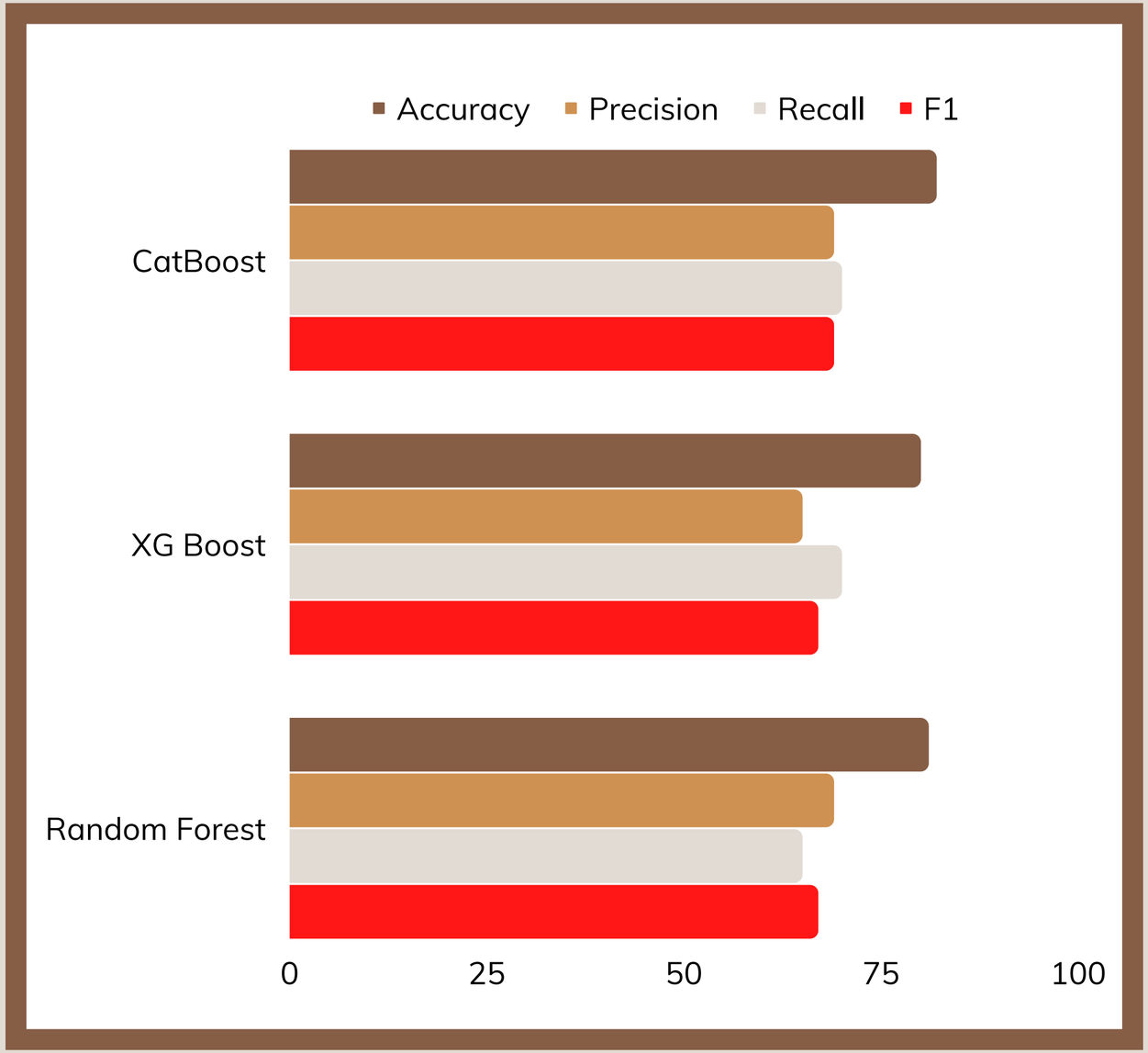
Performance criteria 1: Accuracy

Catboost and Random forest (0,81)

Performance criteria 2: Precision, Recall, F1 Score

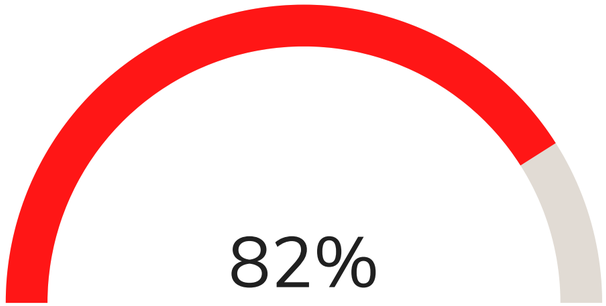
Imbalanced dataset with 0 as a minority class

Cat Boost (0,69 F1score)

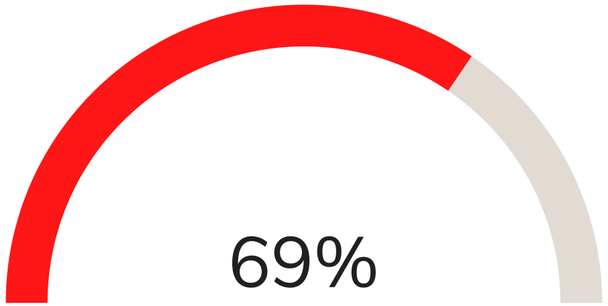


Our Best Predictive Model:

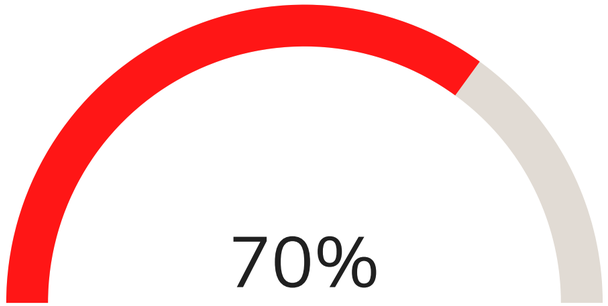
CATBOOST



Accuracy



Precision

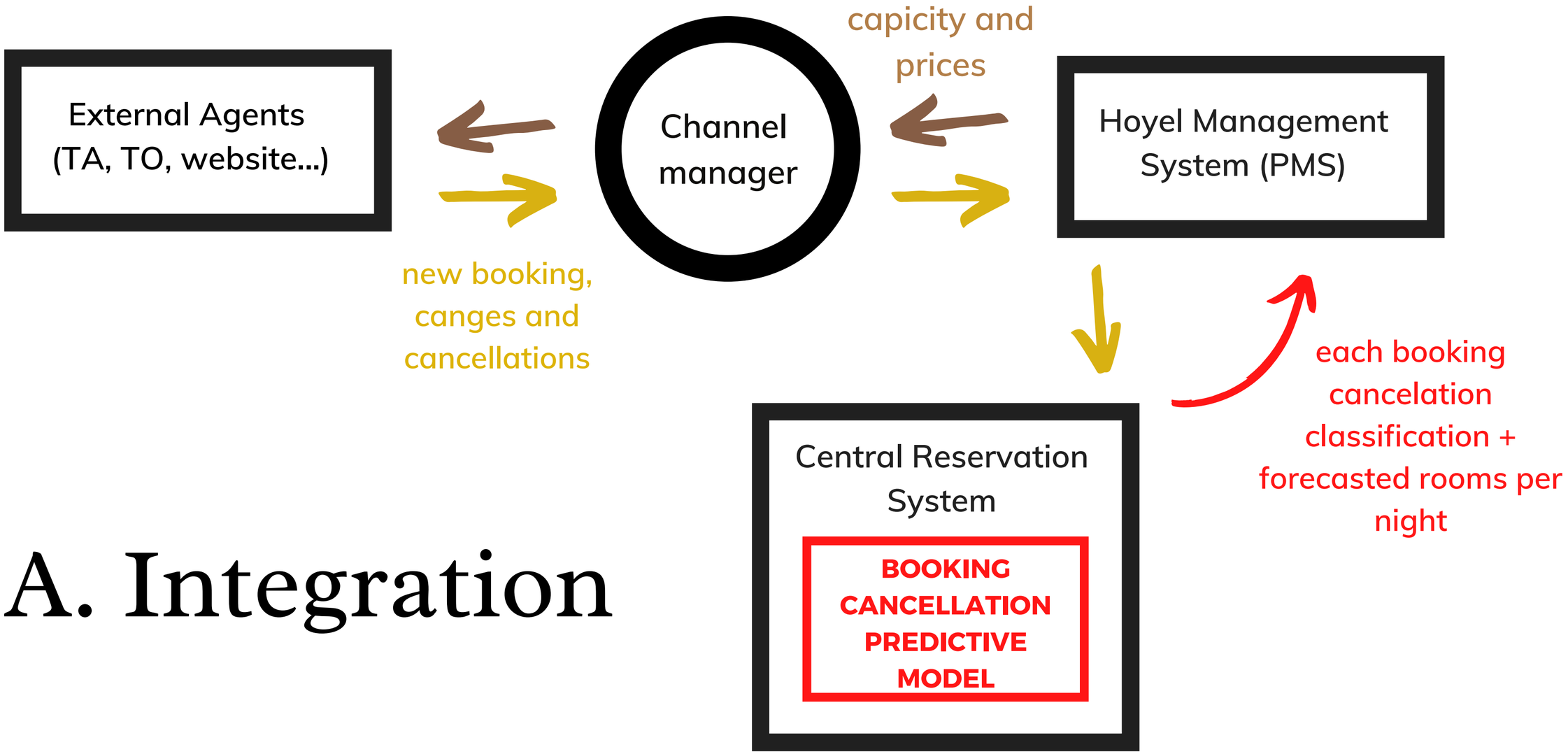


Recall

4. APPLICATION ON BUSINESS

- Integration
- Operations
- Preventive Actions

Integration Operations Preventive Actions



A. Integration

B. Operations

✓ **Acurate forecast of the demands**

Predictive Model run on a daily-basis because of variables' updates

✓ **Enable better management and planning of the channel capacity**

Direct link in real-time between PMS, CRS and External Agents

C. Preventive Actions

Discount on services

access to the spa, gym, rent a car, activities in the surroundings...etc

Special offers

free breakfast, airport pick up...etc

5. CONLUSION

1

CHALLENGES

imbalanced dataset

2

LIMITATION

Data Quality

3

GLOBAL CONCLUSION

THANK YOU FOR ATTENTION! FEEL FREE TO SHARE YOUR QUESTIONS WITH THE TEAM