

PREDICTION OF HOTEL CLUSTER BASED ON HISTORICAL BROWSING DATA

MENTEE: TANZINA ZAMAN

MENTOR: RYAN ROSARIO

DATA SCIENCE INTENSIVE

OVERVIEW

- Introduction
 - Problem description
 - Business need
- Data description
- Exploratory data analysis
- Data analysis
 - Approach 1
 - Approach 2
 - Approach 3
- Result

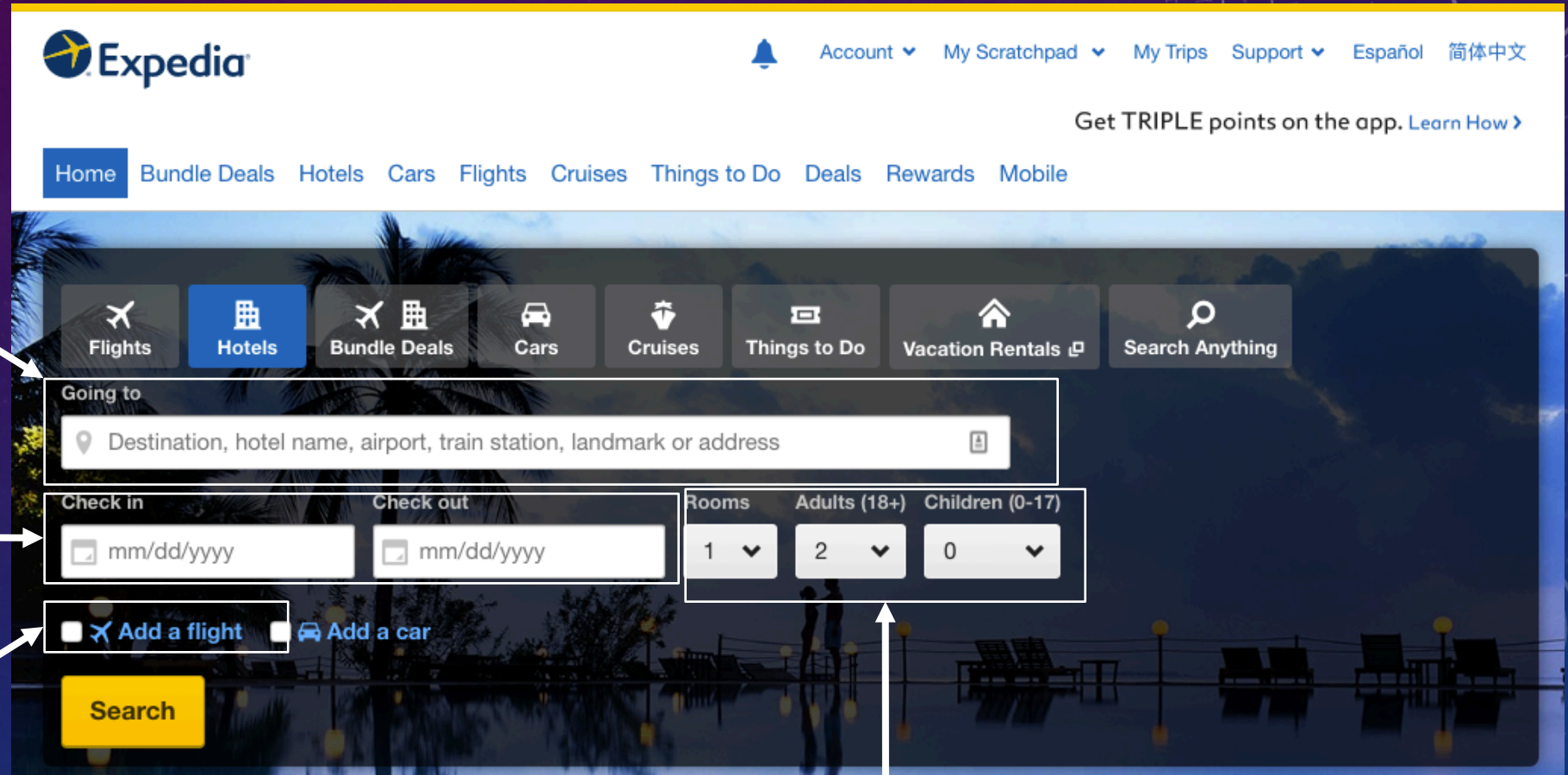
INTRODUCTION

- Problem definition:
 - Predict hotels based on browsing history of customer
- Business need:
 - To help the client to build a recommendation system, which will predict types of hotel the customer is looking for.

DATA DESCRIPTION

- Data source: Kaggle competition sponsored by Expedia
- The training set given in competition is considered as parent source for training and test set for this project.
- The given dataset contains 37 millions observation and 24 fields from year 2013 and 2014
- Aim is to predict the hotel cluster the customer is going to book.

DATA DESCRIPTION



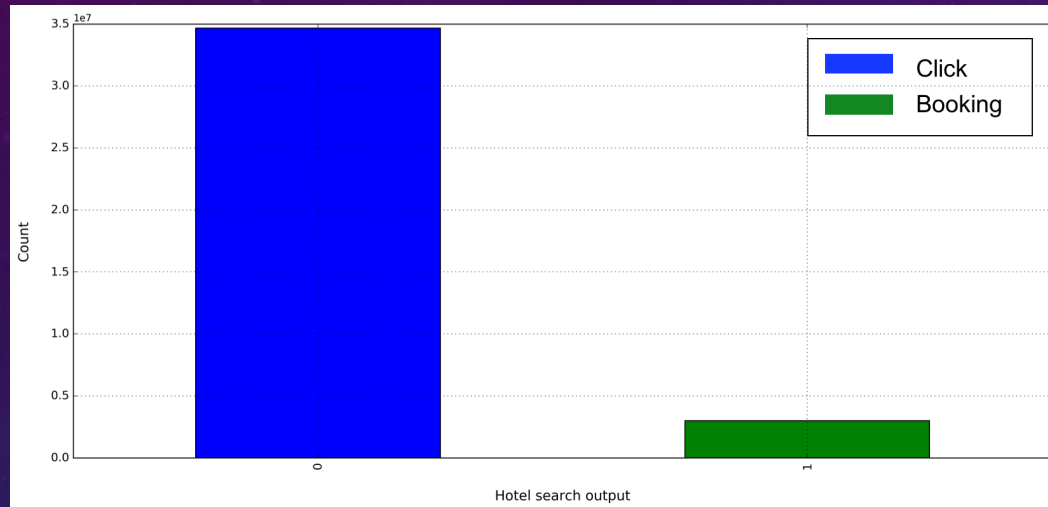
The image shows the Expedia website's search interface with several annotations pointing to specific fields:

- Destination type, continent, country, city etc.** points to the "Going to" input field.
- The desired time for stay by customer** points to the "Check in" and "Check out" date fields.
- Package or not** points to the "Add a flight" and "Add a car" checkboxes.
- Expected occupancy** points to the "Rooms", "Adults (18+)", and "Children (0-17)" dropdown menus.

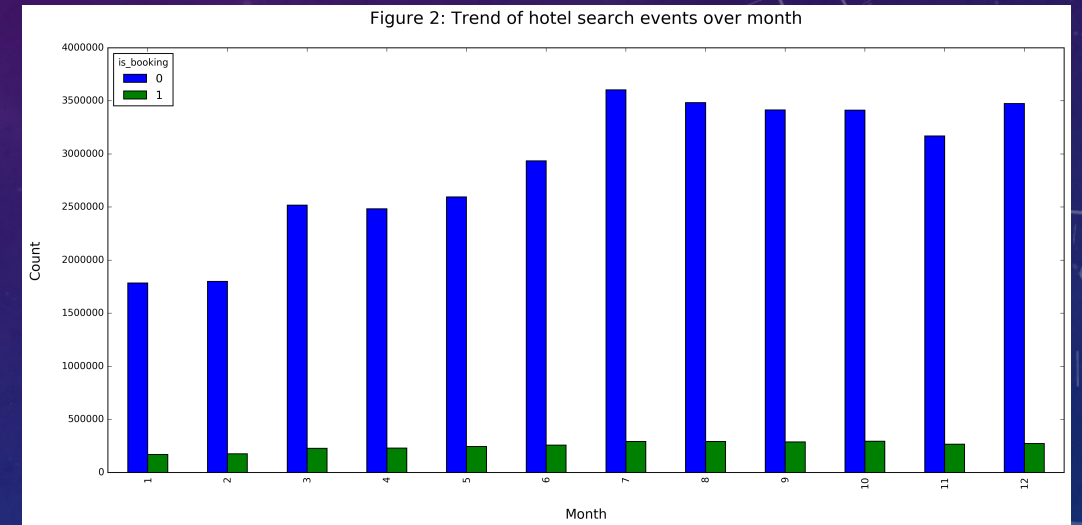
The search interface includes the following elements:

- Header:** Expedia logo, Account, My Scratchpad, My Trips, Support, Español, 简体中文.
- Navigation:** Home, Bundle Deals, Hotels, Cars, Flights, Cruises, Things to Do, Deals, Rewards, Mobile.
- Search Bar:** Flights, Hotels, Bundle Deals, Cars, Cruises, Things to Do, Vacation Rentals, Search Anything.
- Form Fields:**
 - Going to:** Destination, hotel name, airport, train station, landmark or address.
 - Check in:** mm/dd/yyyy
 - Check out:** mm/dd/yyyy
 - Rooms:** 1
 - Adults (18+):** 2
 - Children (0-17):** 0
 - Buttons:** Add a flight, Add a car, Search.

EXPLORATORY DATA ANALYSIS

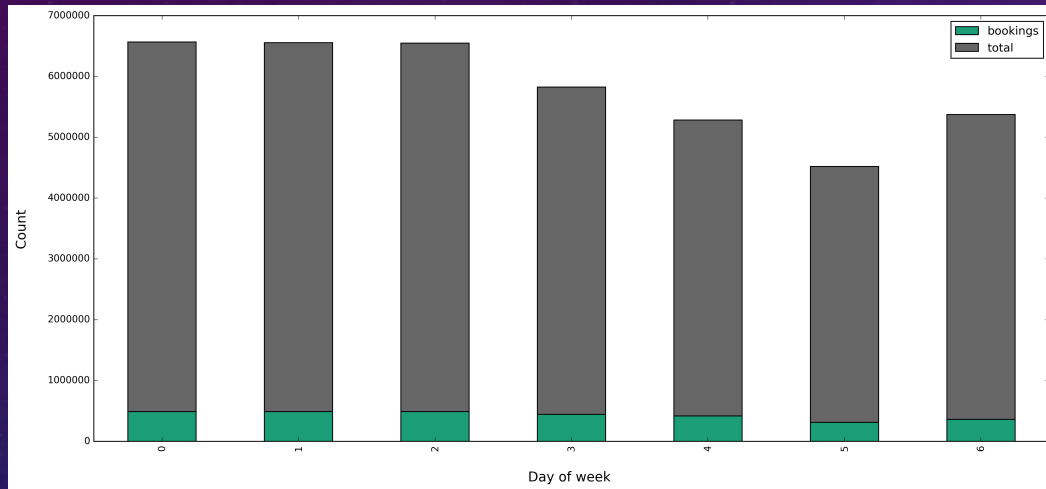


Count of booking and click event present in given dataset

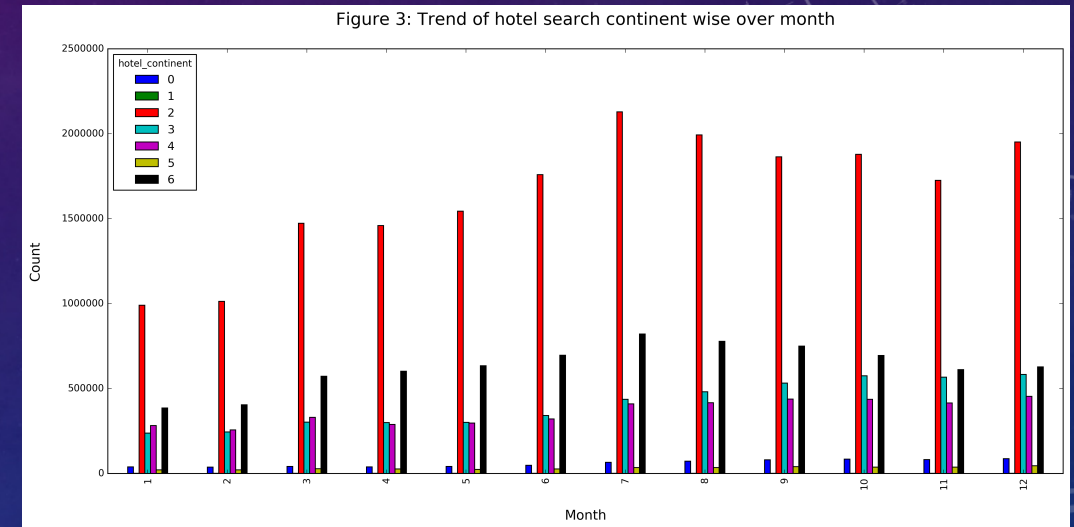


Trend of hotel search over month

EXPLORATORY DATA ANALYSIS

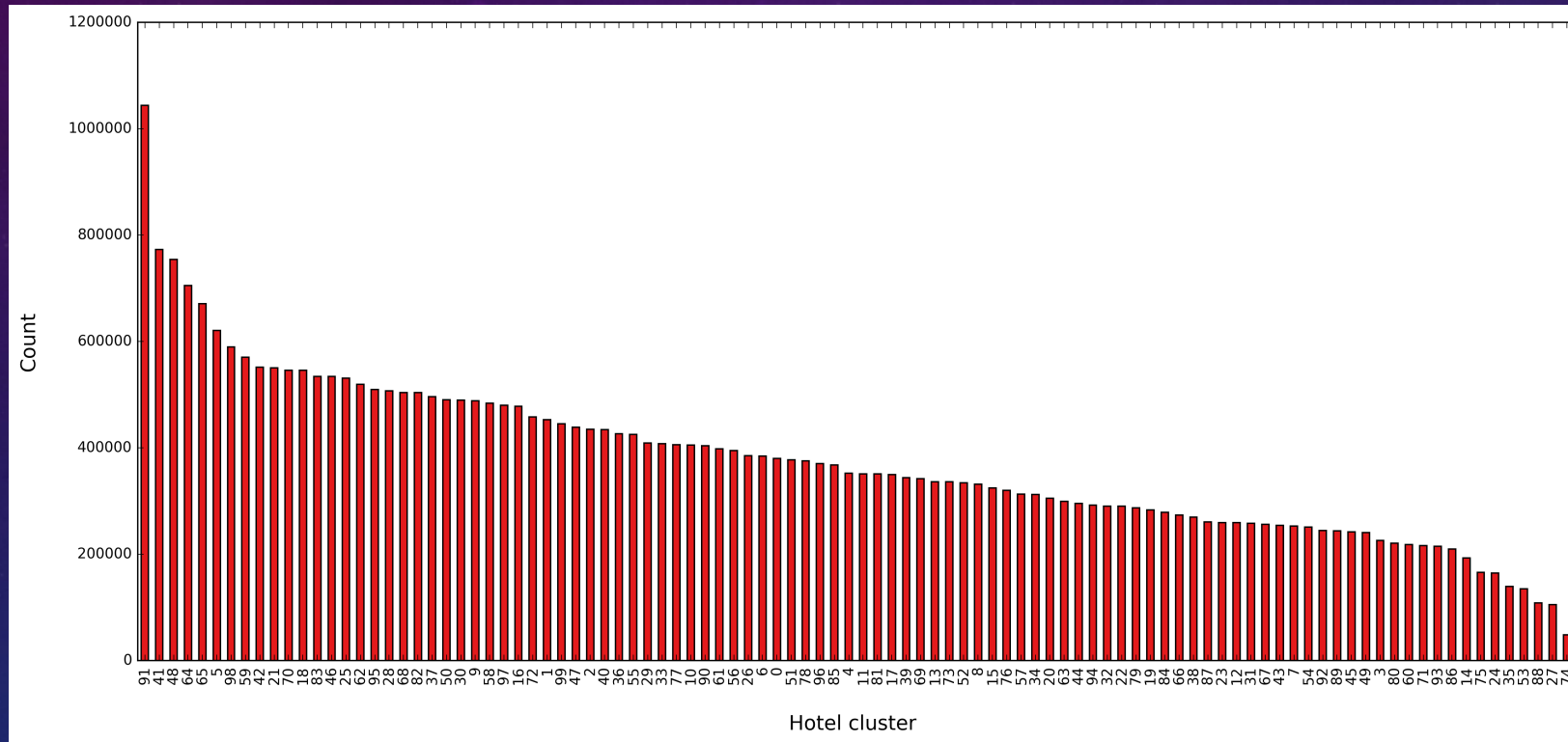


Trend of booking events for days of the week as a ratio of total events



Search trend of hotels in seven continents

EXPLORATORY DATA ANALYSIS



Frequency of hotel cluster

ANALYSIS: APPROACH 1

METHOD

- 1% of given data is randomly sampled and then split into training – test set in 70-30 ratio.
- ID of hotel cluster = Target attribute
- Rest 23 fields = feature
- Correlation among target attribute and feature attribute is calculated
 - No correlation is present; linear regression would not fit here.
- Classification methods (Decision tree, SVM)
 - Unsatisfactory model accuracy; 12% and 3% respectively

ANALYSIS: APPROACH 1

LIMITATION

- Only 1% of 37 million observation is not enough
- The timestamp provided in the given data can be used to create more features
- Total 100 cluster is present as target attribute, which makes hard to predict for machine learning algorithm
- Working with top 10 cluster would make the problem doable.

ANALYSIS: APPROACH 2

METHOD

- Observation for top 10 clusters are selected
- More features are created from the timestamp
- Then split into training – test set.
- 10 Logistic regression model with stochastic gradient descent (SGD) learning is trained as a binary classification.
- Models are evaluated by model accuracy, precision, recall, f1 score, confusion matrix, area under ROC curve.

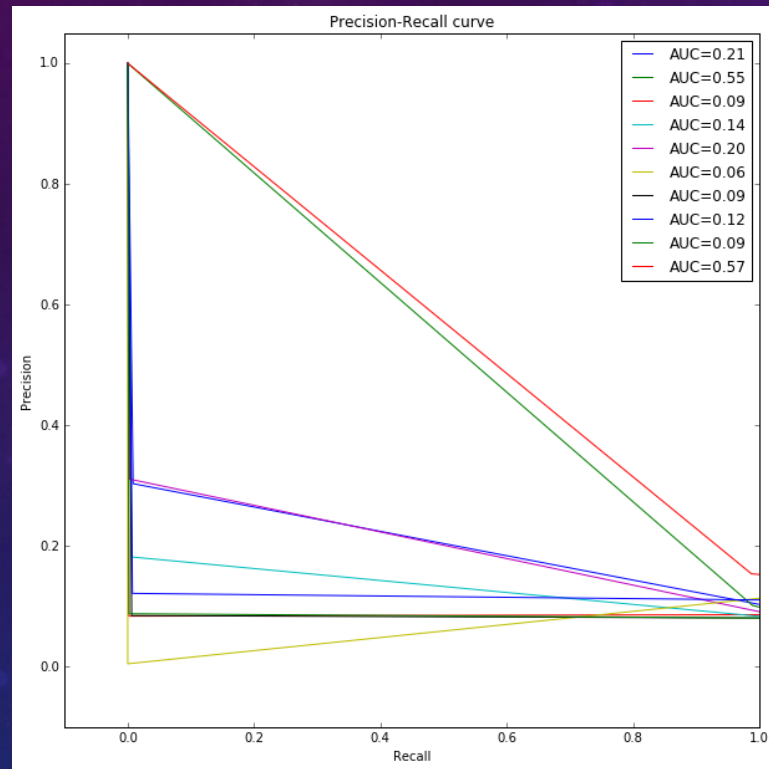
ANALYSIS: APPROACH 2

LIMITATIONS

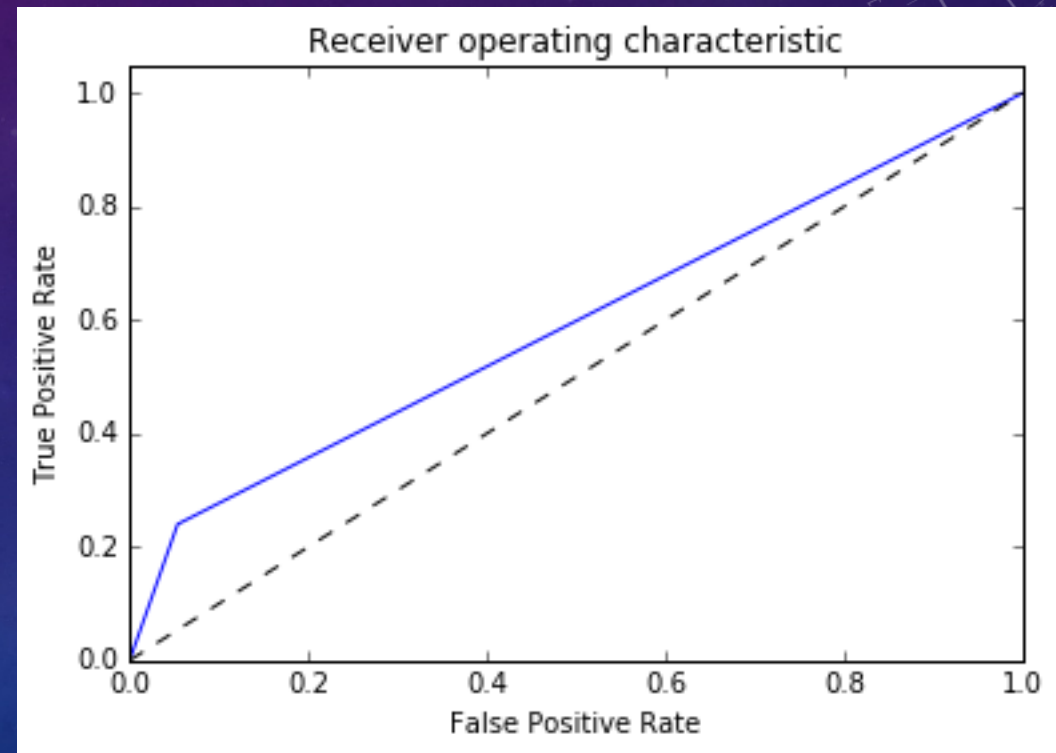
- Model accuracy is good; ranging from 91% to 80%
- Same as precision, recall but area under ROC curve is not up to the mark

Cluster ID	Accuracy %	Precision	Recall	Area under ROC curve
64	87.3	0.91	0.94	0.59
65	89.6	0.90	0.99	0.50
98	91.2	0.91	0.98	0.49
59	91.6	0.91	0.998	0.50
5	90.8	0.90	0.999	0.50
41	88.6	0.88	0.999	0.49
42	90.8	0.91	0.986	0.50
48	88.9	0.88	0.995	0.50
21	91.9	0.91	0.999	0.50
91	82.6	0.84	0.968	0.50

ANALYSIS: APPROACH 2 LIMITATIONS



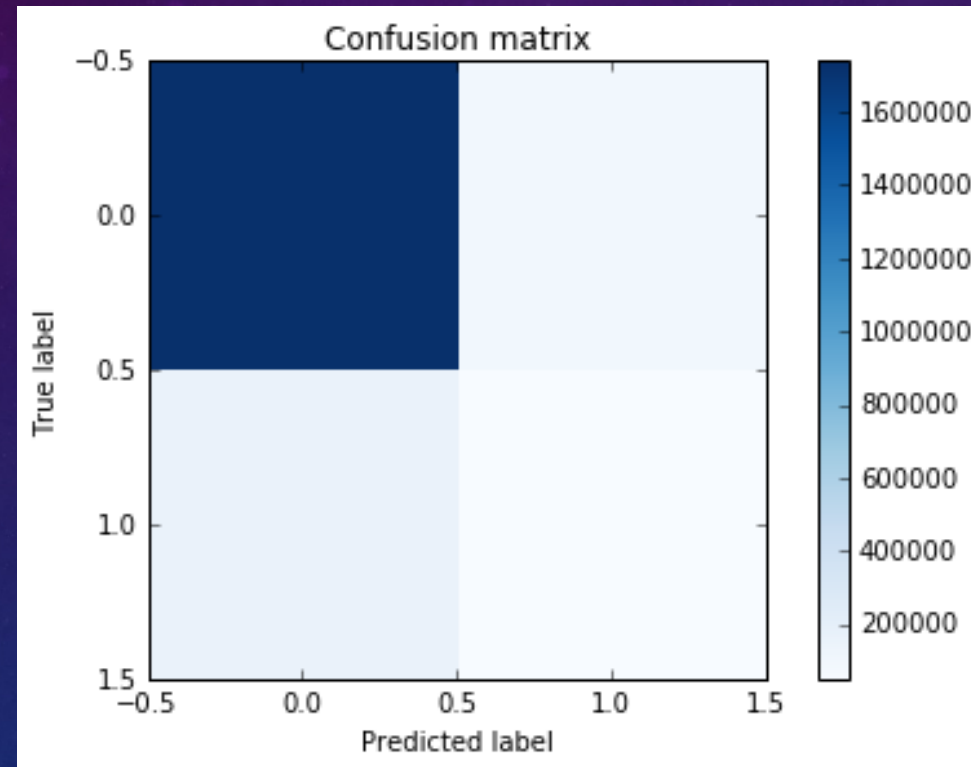
Precision recall curve for all cluster



ROC curve for cluster 64

ANALYSIS: APPROACH 2

LIMITATIONS



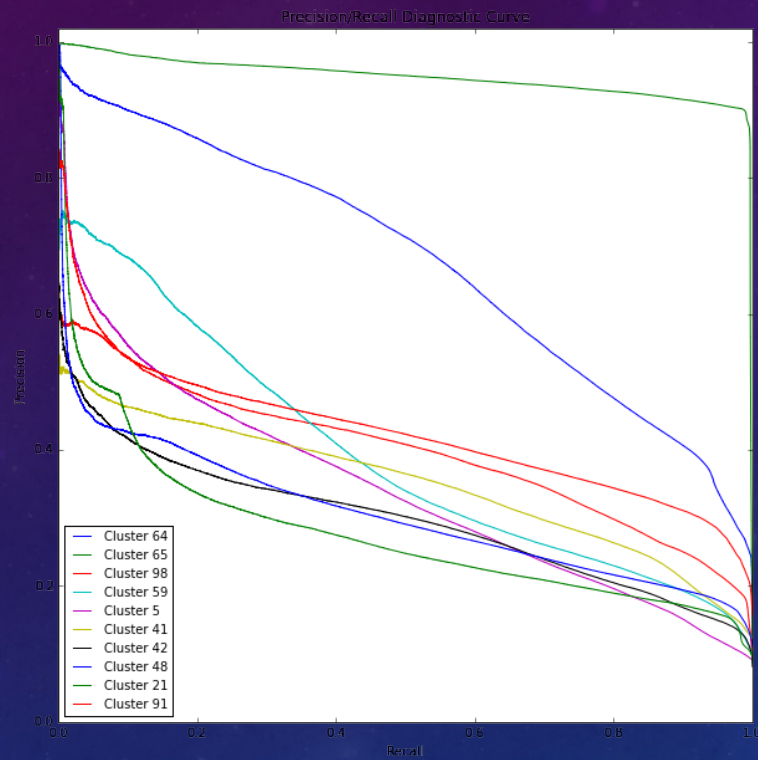
Confusion matrix for cluster 64

ANALYSIS: APPROACH 3

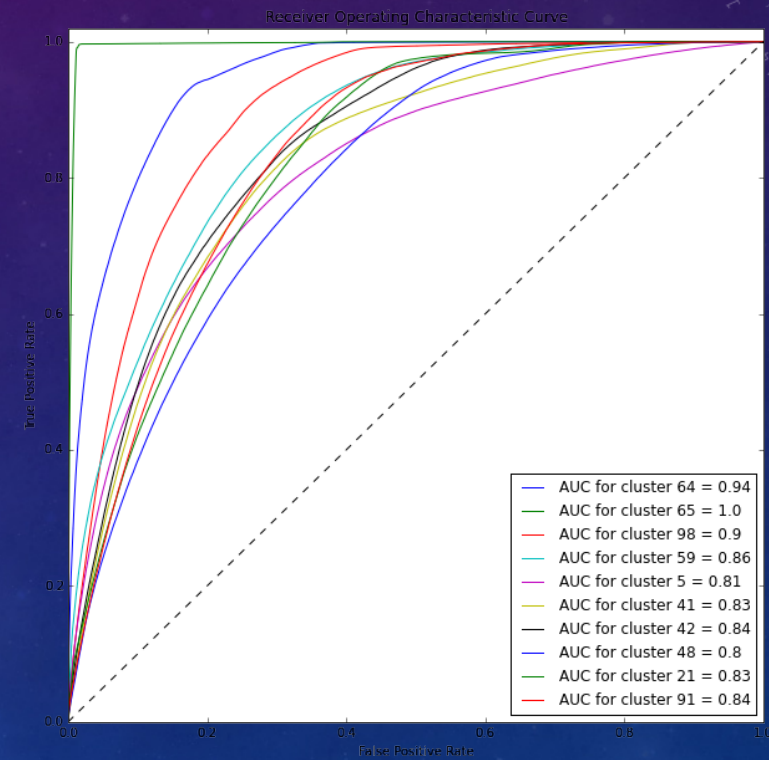
METHOD

- Categorical attributes and numerical attributes are treated separately
- Repeat the training of logistics regression with SGD learning like approach 2
- Models are evaluated using precision-recall curve, area under ROC curve, confusion matrix, precision, recall etc.

RESULT



Precision recall curve for all cluster



ROC curve for all cluster

RESULT

