Yourcabs.com Taxi Cancellation Report

By Abigail McDonald

Introduction

Yourcabs.com, a taxi company in Bangalore, India, was having a problem with drivers canceling calls too close to the scheduled arrival time. This would potentially leave customers without a timely replacement driver. Yourcabs had both an online portal and phone-in booking system. Yourcabs needs to be able to predict which calls will be cancelled.

The dataset contains collected data from 10,000 bookings between the years of 2011 and 2013. Data includes: booking ID, the ID of the customer, vehicle model type, type of package ordered (represents hours and kilometers), type of travel (1=long distance, 2= point to point, 3= hourly rental), unique id for each area (from/to), unique id of city (from/to), time stamp of trip start and end, mobile/desktop booking, time stamp of booking creation, latitude of from/to area, longitude of from/to area, and whether the booking was cancelled (1) or not (0) due to unavailability of a car.

Data Exploration

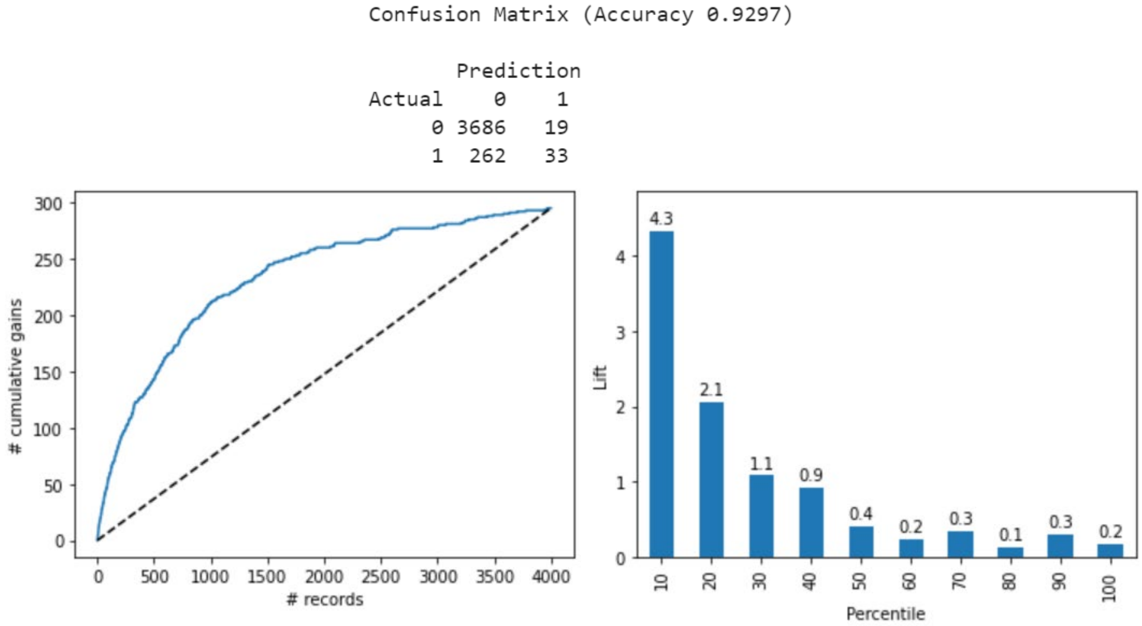
The id and user id columns were dropped as they are personal identifiers and have no influence on the data. The from-city id, to-city id, and to-date were dropped as they contain more nulls than usable values. Additionally, the from/to city information is only applicable to a very small segment of users and the area data better covers user location. The Euclidean distance was calculated using the from/to latitude/longitude columns and made into a new variable ‘distance’. The booking and start date columns were separated into month, day, hour columns in order to use the information. The missing values in the from/to area columns are categorical, and, thus, a new category ‘0’ was added to indicate that the area was not point to point or part of a package. Additionally, any packages in the package id column that were null were made into ‘0’ packages. Categorical data (from area, to area, and vehicle model) were dropped when using models that cannot run with categorical data. Travel type and package id were one-hot encoded to become useful in this situation.

Booking appears to be done in the same month and, usually, the same day (>75%). Phone call booking appears to be the most popular method, followed by online and then mobile site booking. Late summer/early fall is the most popular time of year to book, although month and day of week are mostly evenly distributed. There are spikes in booking around start of work (am) and end of work (pm). Booking for month/day seem to follow the start date density trends. Finally, more bookings seem to be made in the morning for services.

Data Analysis

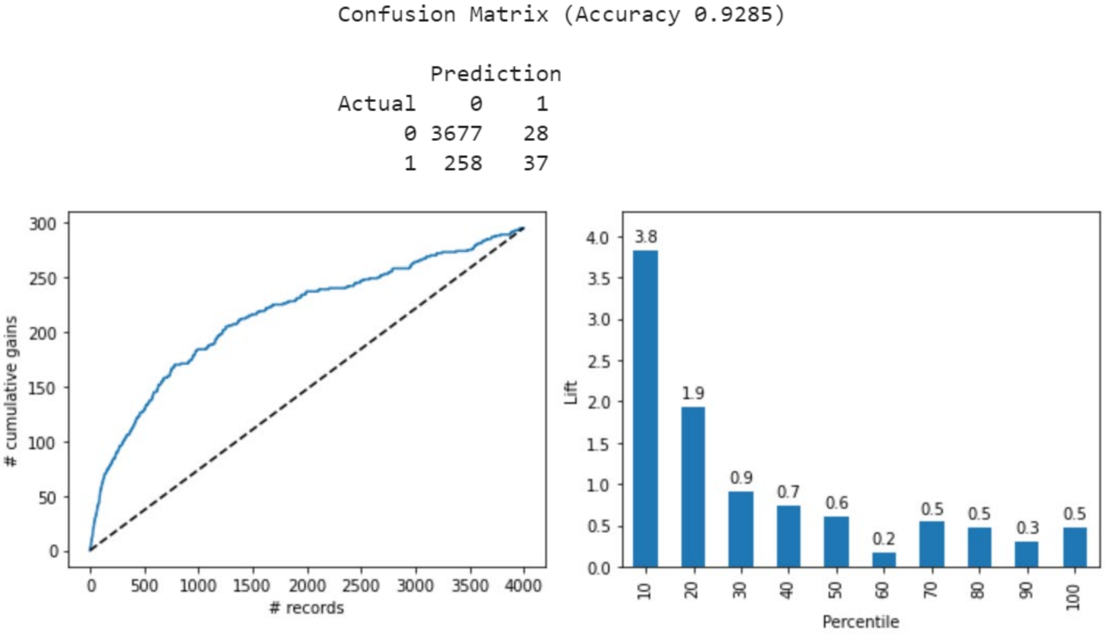
The best performing predictive models were the Random Forest model and the KNN model at 92.97% and 92.85% respectively. Additionally, their cumulative gains charts showed nice, rounded lifts in the middle of the data. Also, both models performed exceptionally well in the first decile. This demonstrates that they both have considerable and reliable predictive power when compared to other models used.

Confusion Matrix, Gains, and Lift Charts for Random Forest Model



```
                Confusion Matrix (Accuracy 0.9297)

                   Prediction
                Actual    0     1
                     0 3686    19
                     1  262    33
```

Confusion Matrix, Gains, and Lift Charts for KNN Model



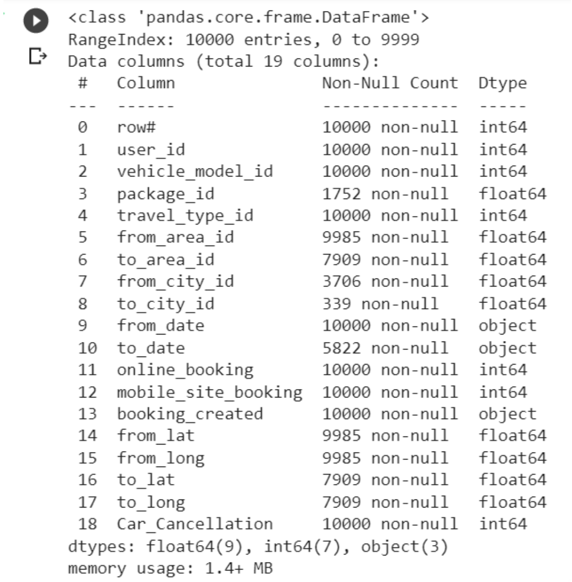
```
                Confusion Matrix (Accuracy 0.9285)

                   Prediction
                Actual    0     1
                     0 3677    28
                     1  258    37
```

Conclusion

Five models were considered in predicting cancellations from the Yourcabs.com taxi data: a deep classification tree, a Random Forest model, Linear Discriminant Analysis, a KNN model, and a Naïve Bayes model. The Random Forest model performed the best and the KNN model a close second in predictive ability. When looking for important features, it was determined that distance was negatively correlated with car cancellation (the further the distance, the more likely to be cancelled). Distance was determined to be the most important feature impacting cancellation (Figure 3) followed by to/from area and to/from hour (time of day). Suggestions for future determination of cancellations would be to use the random forest model and to flag features matching high cancellation scenarios.

Appendix

Table A - Null Values in the Dataset



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 19 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   row#                 10000 non-null  int64
 1   user_id              10000 non-null  int64
 2   vehicle_model_id     10000 non-null  int64
 3   package_id           1752 non-null   float64
 4   travel_type_id       10000 non-null  int64
 5   from_area_id         9985 non-null   float64
 6   to_area_id           7909 non-null   float64
 7   from_city_id         3706 non-null   float64
 8   to_city_id           339 non-null    float64
 9   from_date            10000 non-null  object
 10  to_date              5822 non-null   object
 11  online_booking       10000 non-null  int64
 12  mobile_site_booking  10000 non-null  int64
 13  booking_created      10000 non-null  object
 14  from_lat             9985 non-null   float64
 15  from_long            9985 non-null   float64
 16  to_lat               7909 non-null   float64
 17  to_long              7909 non-null   float64
 18  Car_Cancellation     10000 non-null  int64
dtypes: float64(9), int64(7), object(3)
memory usage: 1.4+ MB
```

Figure 1 - Correlation of Variables

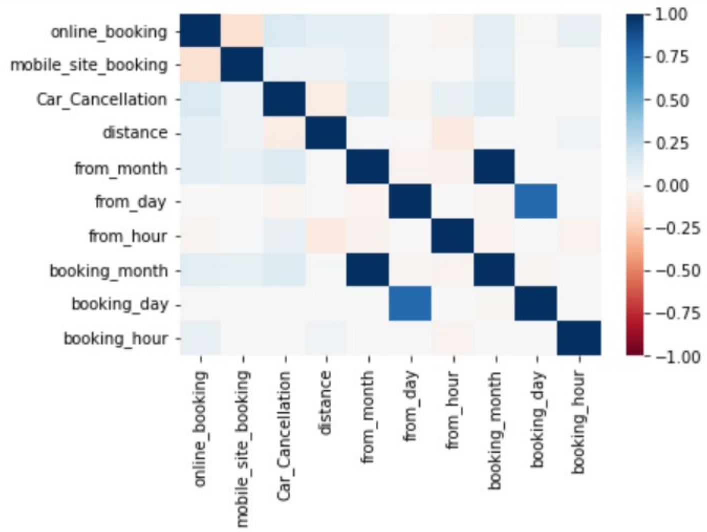


Figure 2 - Density of Variables

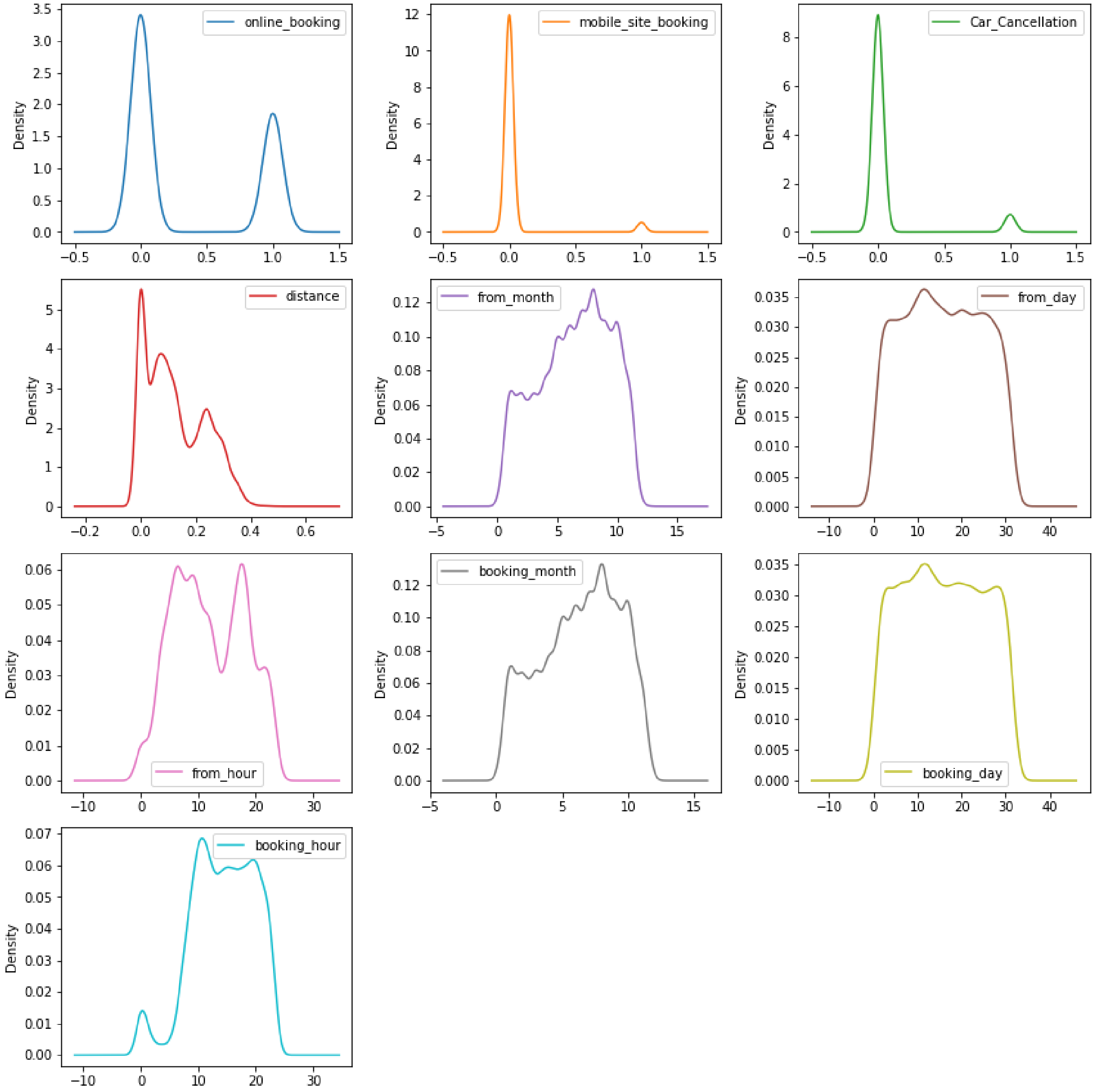
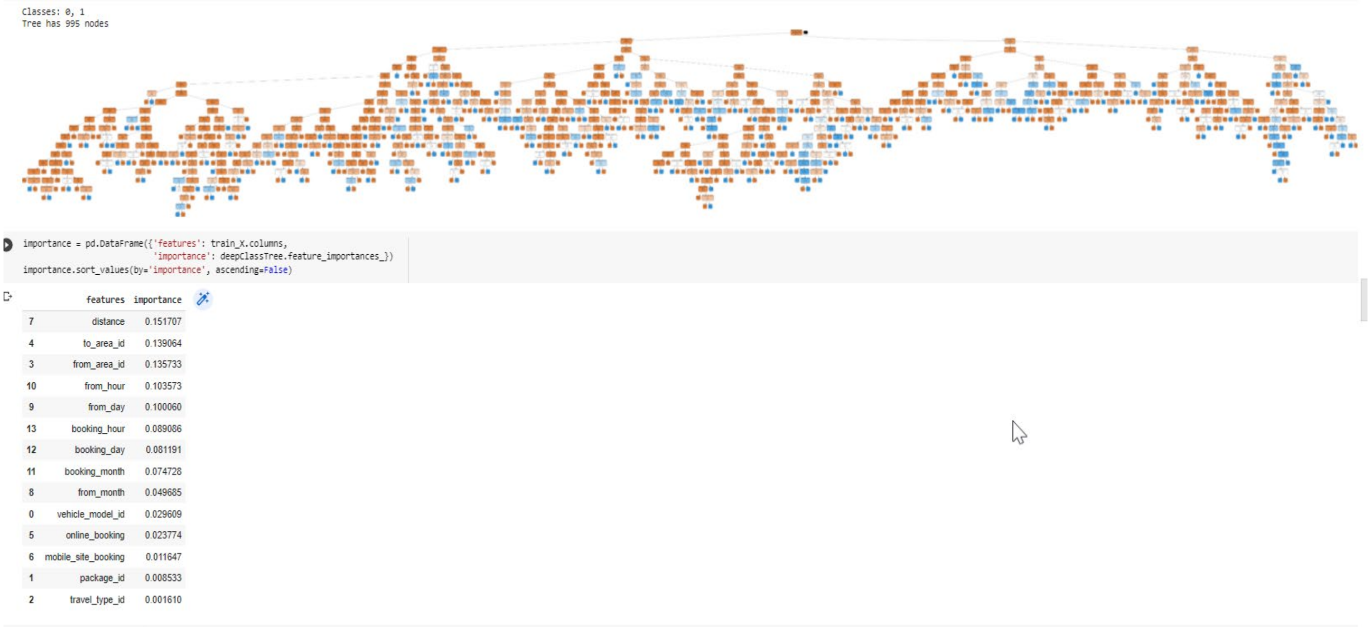


Figure 3 - Classification Tree with Feature Importance



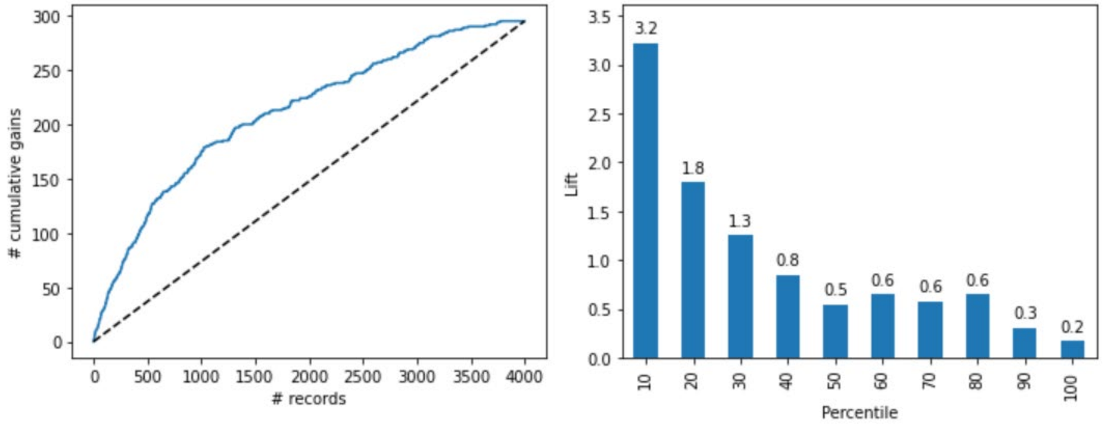
```
Classes: 0, 1
Tree has 995 nodes
```

```
importance = pd.DataFrame({'features': train_X.columns,
                           'importance': deepClassTree.feature_importances_})
importance.sort_values(by='importance', ascending=False)
```

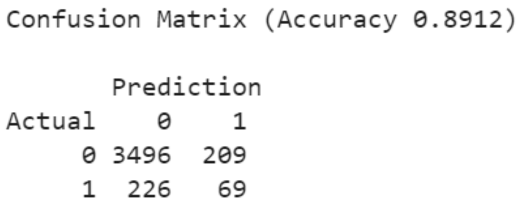
	features	importance
7	distance	0.151707
4	to_area_id	0.139064
3	from_area_id	0.135733
10	from_hour	0.103573
9	from_day	0.100060
13	booking_hour	0.089086
12	booking_day	0.081191
11	booking_month	0.074728
8	from_month	0.049685
0	vehicle_model_id	0.029609
5	online_booking	0.023774
6	mobile_site_booking	0.011647
1	package_id	0.008533
2	travel_type_id	0.001610

Confusion Matrix, Gains, and Lift Charts for Linear Discriminant Analysis

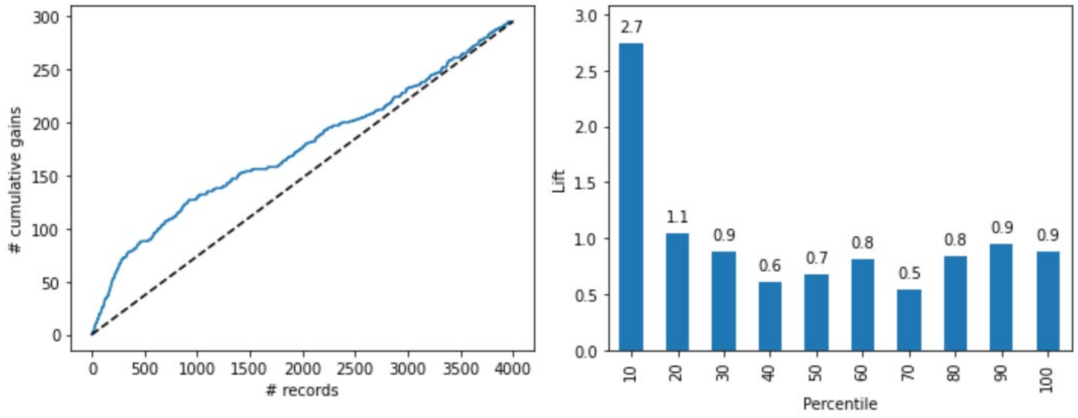
```
Confusion Matrix (Accuracy 0.9263)

       Prediction
Actual    0    1
     0 3703    2
     1  293    2
```



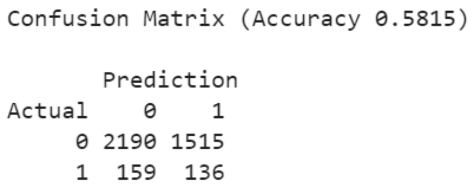
Confusion Matrix, Gains, and Lift Charts for Classification Tree Model



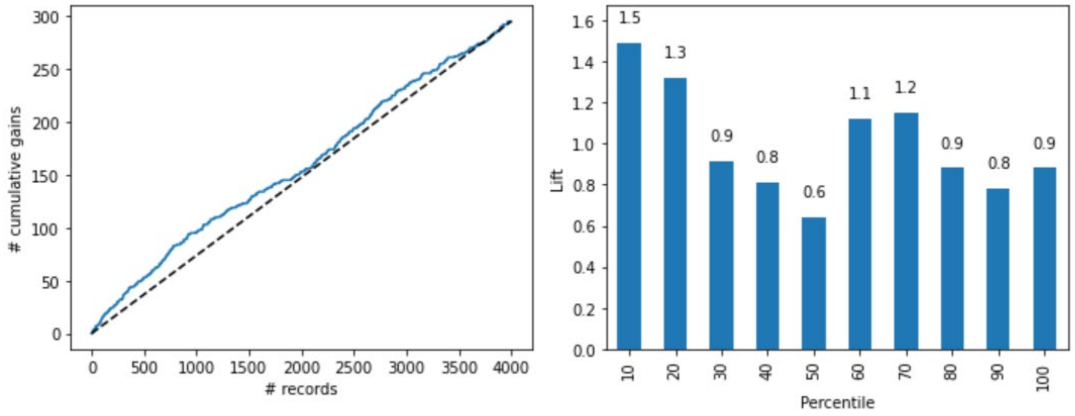
```
Confusion Matrix (Accuracy 0.8912)

       Prediction
Actual    0    1
     0 3496  209
     1  226   69
```



Confusion Matrix, Gains, and Lift Charts for Naïve Bayes Model



```
Confusion Matrix (Accuracy 0.5815)

       Prediction
Actual    0    1
     0 2190 1515
     1  159  136
```


# cumulative gains
# records
Lift
Percentile