



預測模型-分類模型

Term deposit account opening forecasting
based on marketing campaign info.

(透過行銷活動資訊預測銀行客戶是否會開設定期存款帳戶)

1. Industry Needs

- Identify potential customer of term deposit account
- Realize precision marketing

2. Objective

Develop an accurate automation forecasting model of identifying who will open a term deposit account based on marketing campaign (Telemarketing) information.

3. The Proposed Model

- Utilize Gradient Boosting Classifier to forecast potential customers.
- Adjust classification model threshold to get better results, especially in recall rate.

4. Performance

- Precision: 84.73%, Recall: 85.55%
- F1 Score: 85.14%
- AUC: training data achieves 0.86 and testing data achieves 0.82, which is excellent discrimination (80%-90%).

5. Findings

- The proposed classification model has excellent discrimination in identifying potential customer.
- Features of potential customer of term deposit account:
 1. Customers who talked longer than 7.1 minutes
 2. People with balance > \$1,708
 3. Students and retired people

Methodology

Abstract

Data
Understanding

Data
Preparation

Modeling

Results

Data Understanding

- Used Data: Information of 11,162 bank clients (row)
- Feature/Column: x (16 features) + y (has the client subscribed a term deposit? 'yes','no')

Bank Client Data		Campaign Data	
age	age	contact	contact communication type ('cellular','telephone')
job	type of job('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')	month	last contact month of year
marital	marital status('divorced','married','single','unknown')	day	day number of last contact day
education	primary, secondary, tertiary and unknown	duration	last contact duration, in seconds
default	has credit in default?('no','yes','unknown')	campaign	number of contacts performed during this campaign and for this client
housing	has housing loan?('no','yes','unknown')	pdays	number of days that passed by after the client was last contacted from a previous campaign
loan	has personal loan?('no','yes','unknown')	previous	number of contacts performed before this campaign and for this client
balance	Balance of the individual	poutcome	outcome of the previous marketing campaign('failure','nonexistent','success')

Methodology

Abstract

Data
Understanding

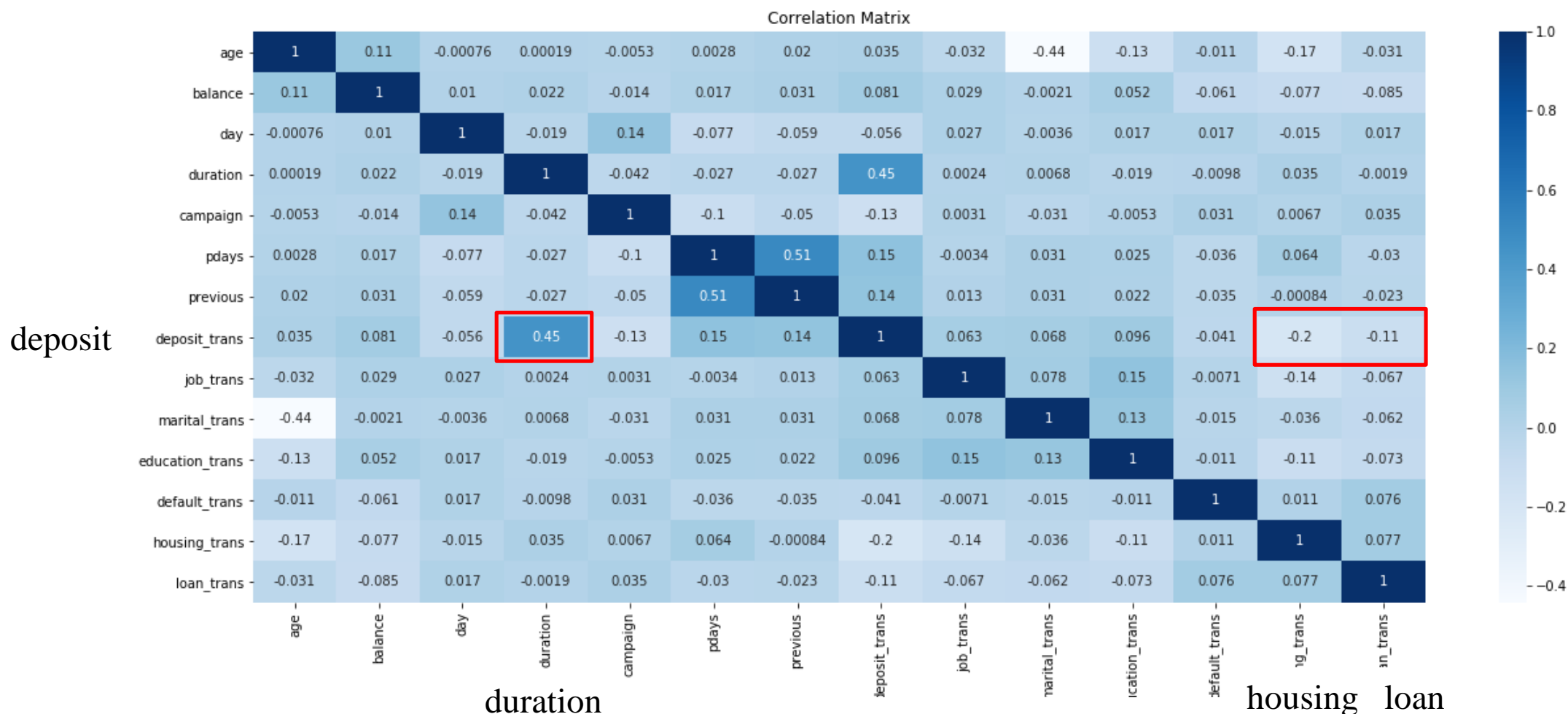
Data
Preparation

Modeling

Results

Data Understanding - Correlation Matrix

- There is no correlation coefficient achieves $\pm 70\%$ (strong correlated)
- Correlation coefficient between deposit and duration reaches 45%



Methodology

Abstract

Data
Understanding

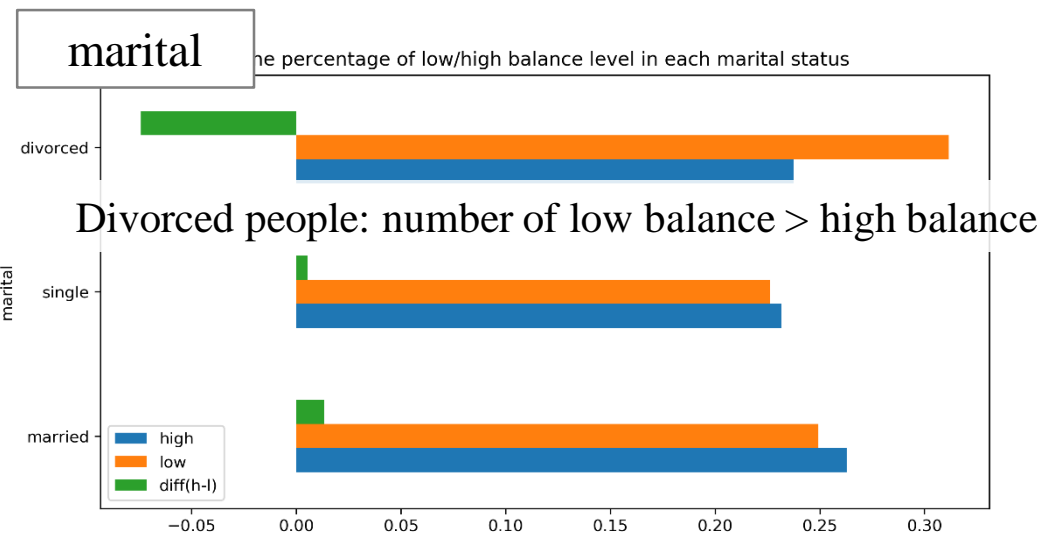
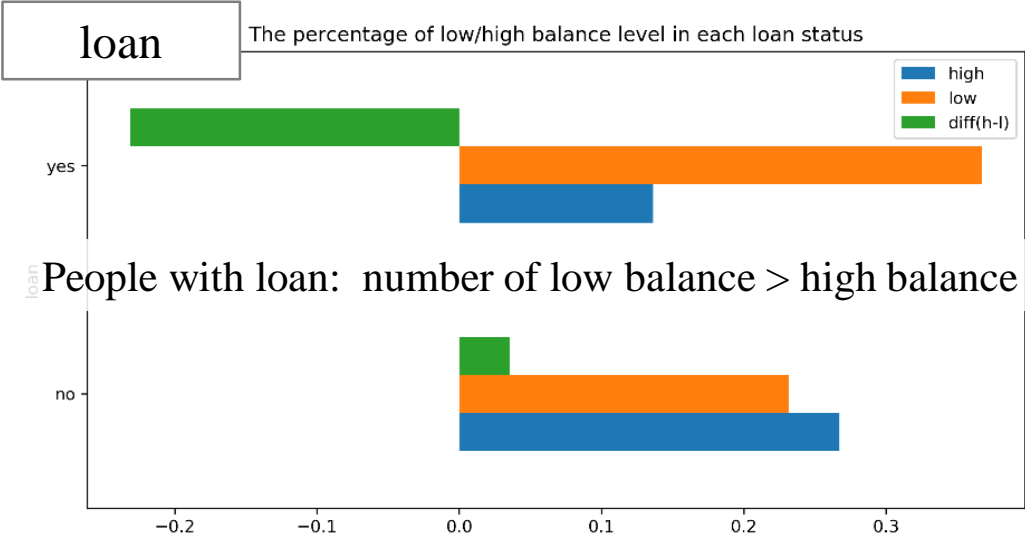
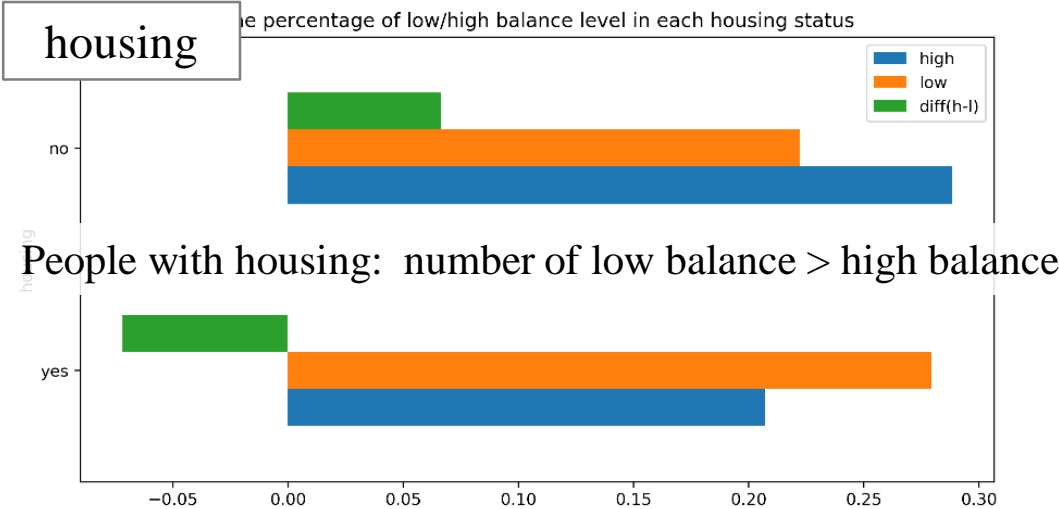
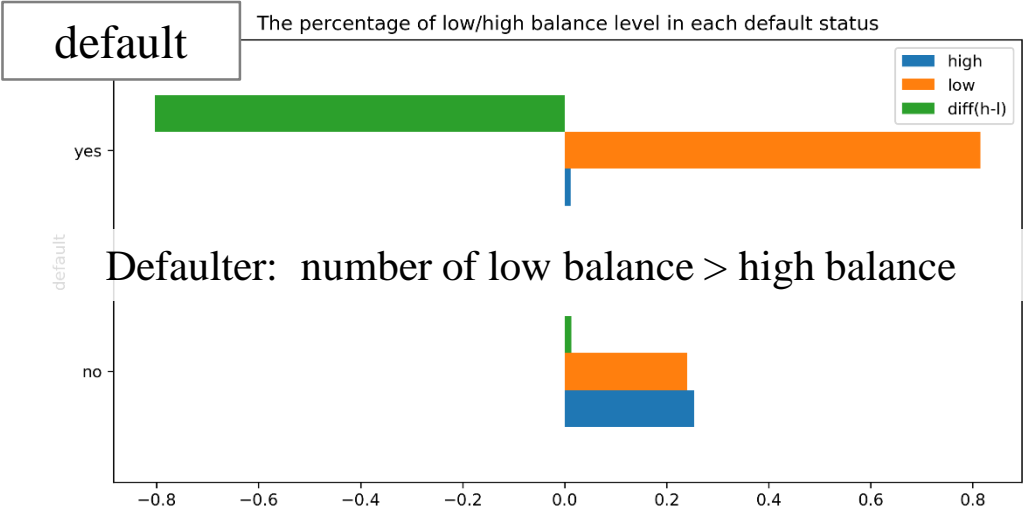
Data
Preparation

Modeling

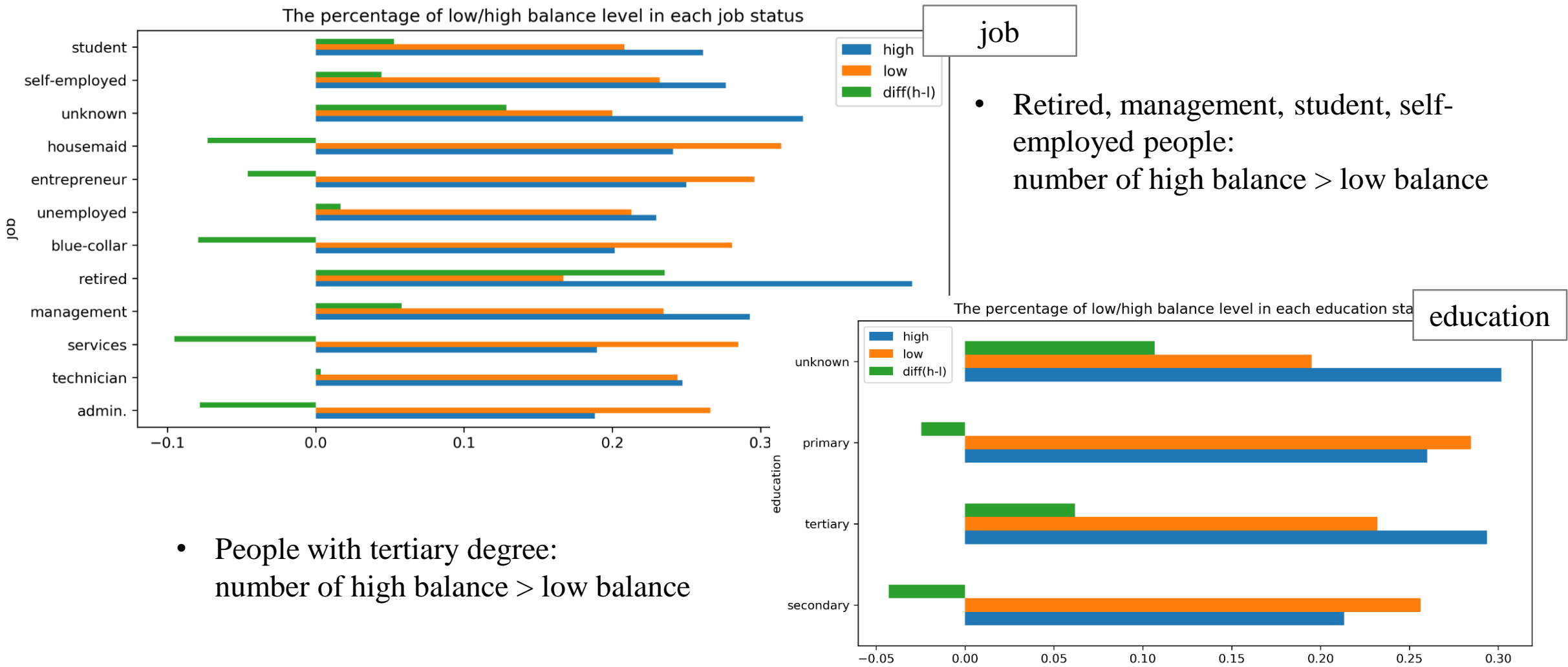
Results

Data Understanding - Relationship between balance status and other client basic info.

- balance status : **high**(>75%), **low**(<25%), **diff**(high-low)



Data Understanding - Relationship between balance status and other client basic info.



Methodology

Abstract

Data
Understanding

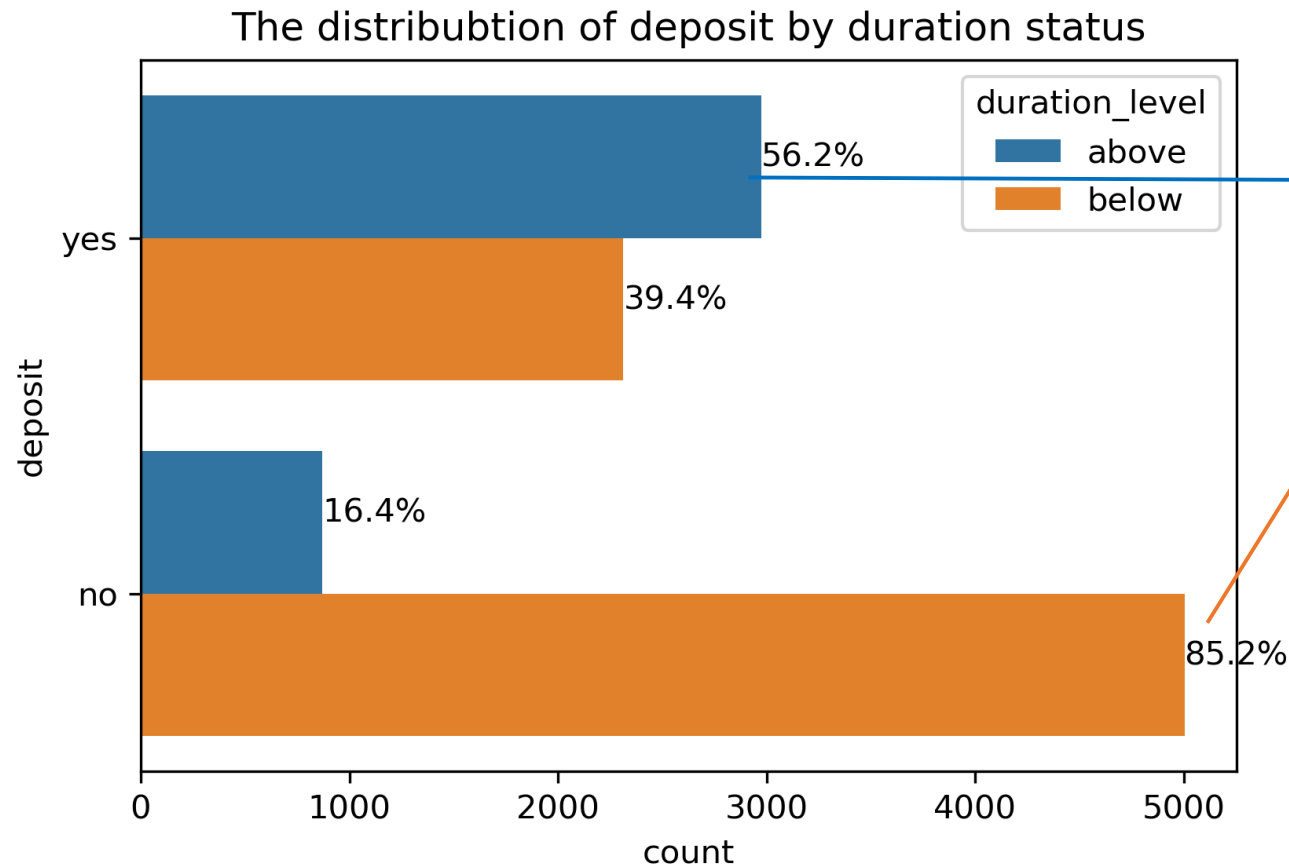
Data
Preparation

Modeling

Results

Data Understanding - deposit vs. duration status

- duration_level: **above**(higher than the average), **below**(lower than the average)



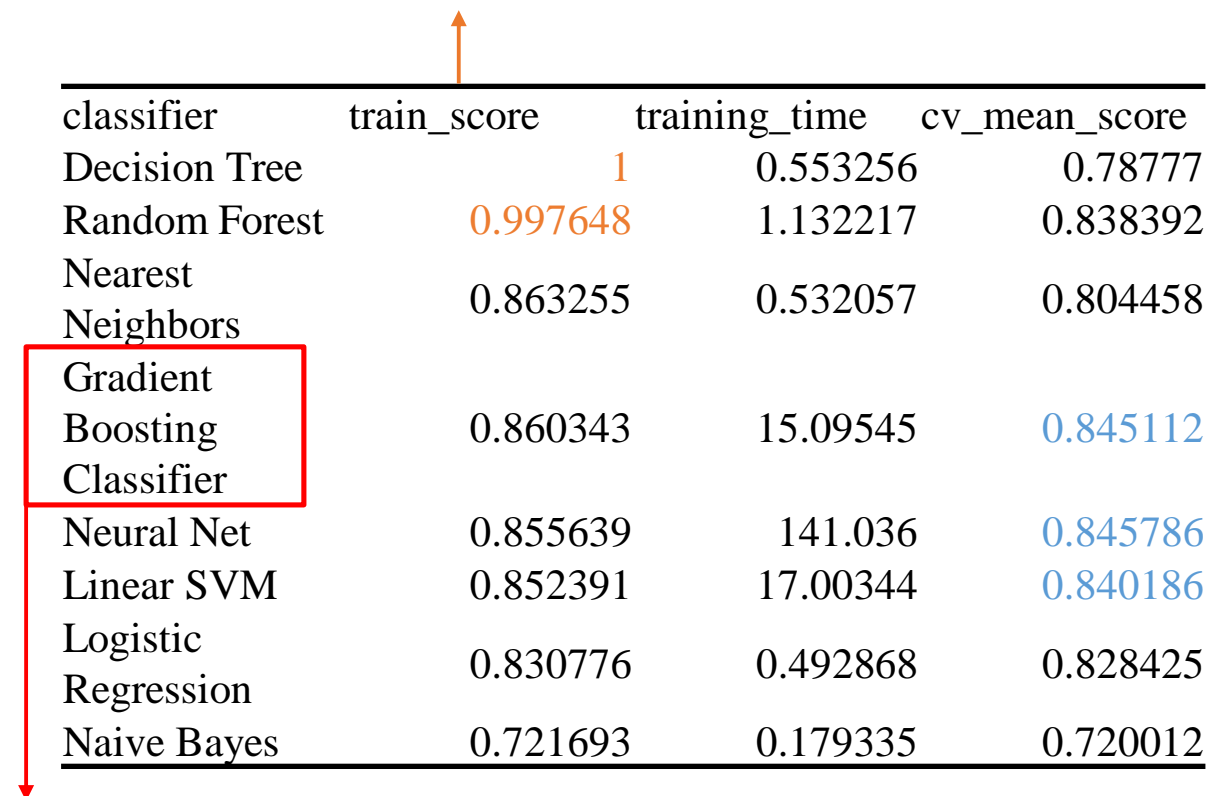
- 56.2% of the people who opened the account talked longer than the average
- 85.2% of the people who did not open the account talked shorter than the average

Data Preparation

Data Processing	Variable	Content
Label Encoding/ One-Hot Encoding	All categorical variables	Encode categorical features as a one-hot numeric array. (LabelEncoder transforms y into 0 and 1, and Onehotencoder is applied to x variables)
Cross –Validation (StratifiedShuffleSplit)	All variables	Validate the model through generating different combinations of the data we already have. (Stratified based on feature 'loan' due to it's relative high correlation coefficient and high uneven distribution degree of y)
Feature Standardization (StandardScaler)	All numeric variables of x	Make distribution of the data have a mean value 0 and standard deviation of 1 (Standard Normal Distribution).

Training Score and Cross-validation Score

1. Train_score (MAE, Mean Absolute Error): Top 2 (**Decision Tree and Random Forest**) may be overfitting



classifier	train_score	training_time	cv_mean_score
Decision Tree	1	0.553256	0.78777
Random Forest	0.997648	1.132217	0.838392
Nearest Neighbors	0.863255	0.532057	0.804458
Gradient Boosting Classifier	0.860343	15.09545	0.845112
Neural Net	0.855639	141.036	0.845786
Linear SVM	0.852391	17.00344	0.840186
Logistic Regression	0.830776	0.492868	0.828425
Naive Bayes	0.721693	0.179335	0.720012

2. Cv_mean_score (cv=3):
Top 3 (**Neural Net, Gradient Boosting Classifier, and Linear SVM**)

3. In top 3 of cv_mean_score, **Gradient Boosting Classifier** has the highest training score

Modeling

Abstract

Data
Understanding

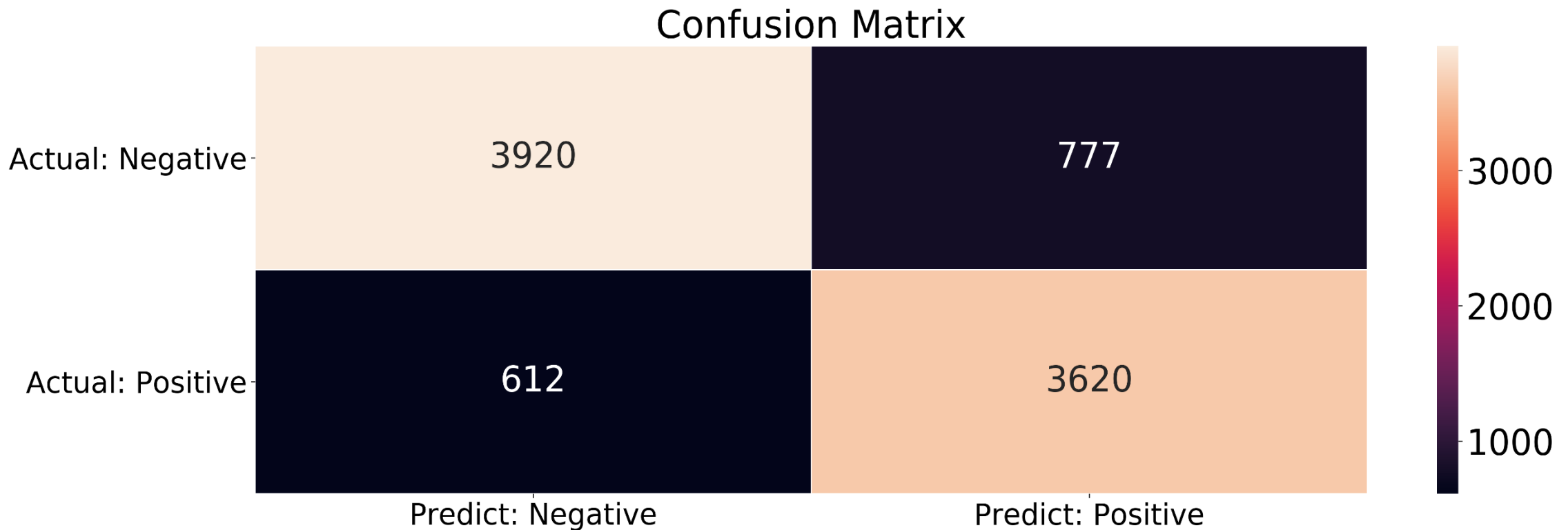
Data
Preparation

Modeling

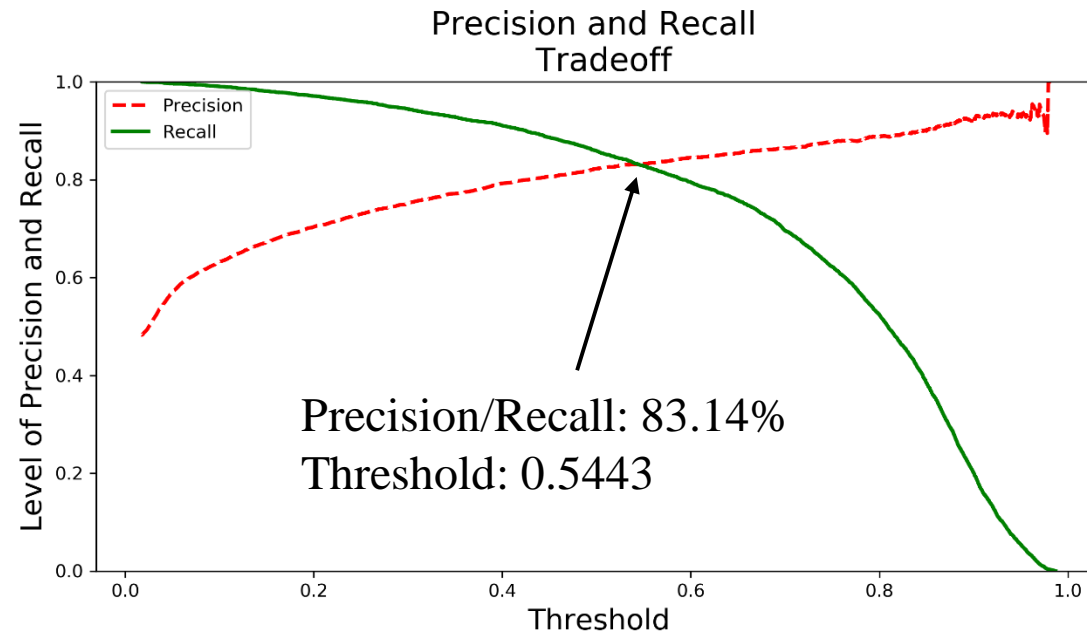
Results

Confusion Matrix - Gradient Boosting Classifier

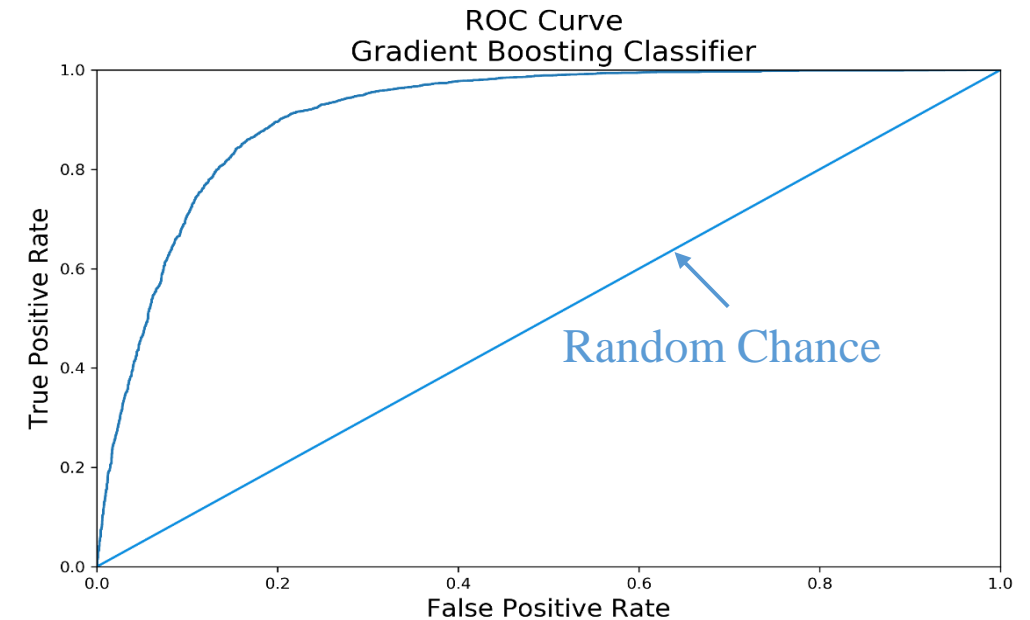
- Precision: 82.44%, Recall: 85.52%, F1 Score: 83.95%
- When this model says these people will open account, 82.44% of the time is accurate.
- 85.52% of the people who will open account is detected by this model.



ROC Curve - Gradient Boosting Classifier



When threshold is 0.5443, precision is equal to recall 83.14%. To get higher recall and precision, new threshold is set up to predict again.



Under default threshold (0.5), AUC reaches 0.9121

Results

Abstract

Data
Understanding

Data
Preparation

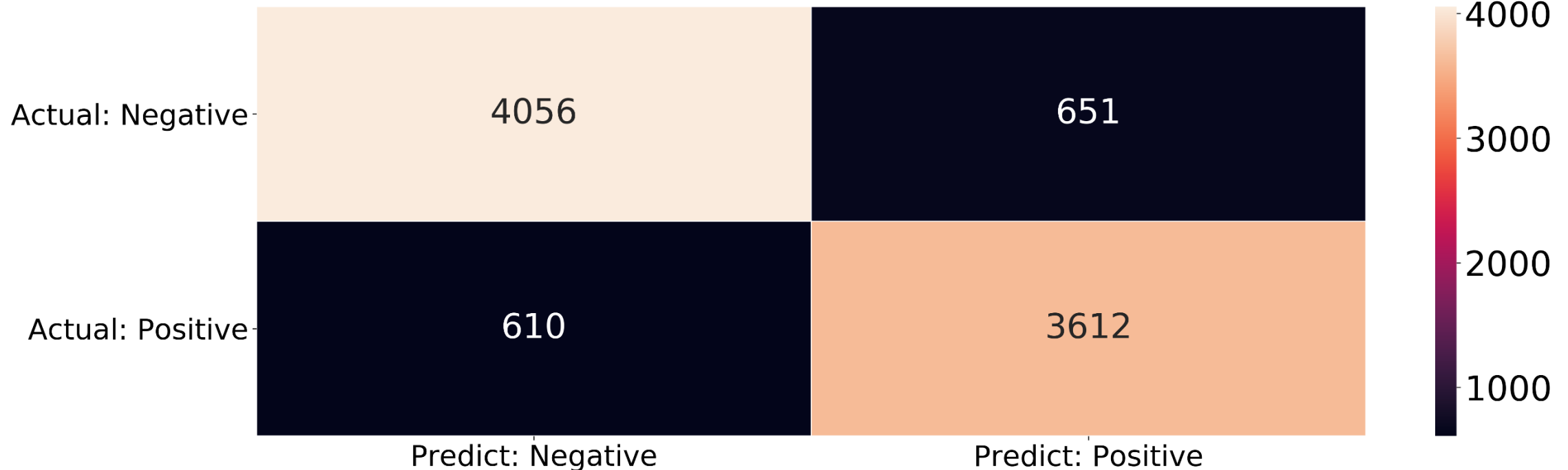
Modeling

Results

Confusion Matrix - Gradient Boosting Classifier (new threshold 0.5443)

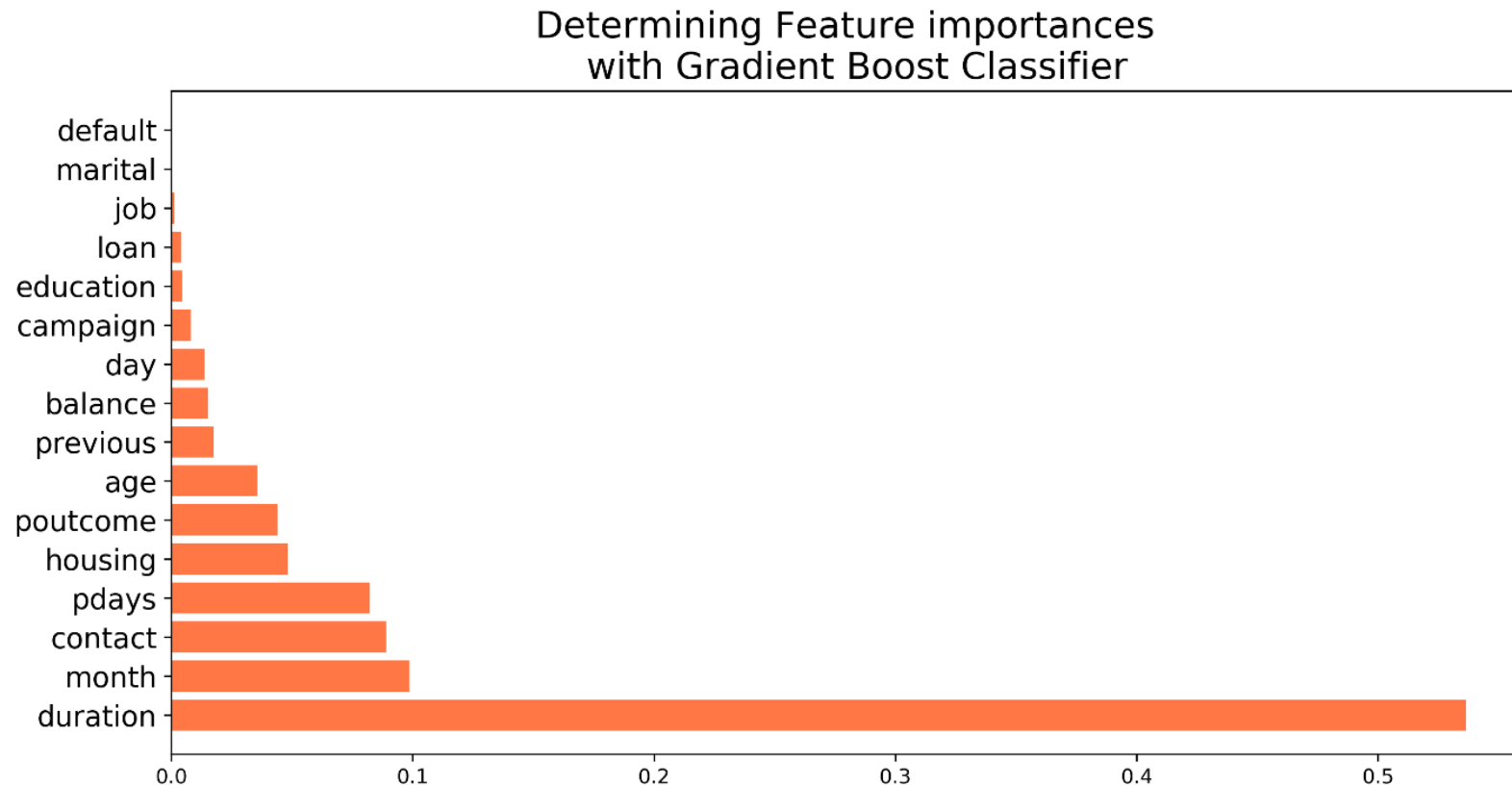
- Precision: 84.73%, Recall: 85.55%, F1 Score: 85.14%
- AUC: training data achieves 0.86 and testing data achieves 82%, which is excellent discrimination (80%-90%).
- When this model says these people will open account, 84.73% of the time is accurate.
- 85.55% of the people who will open account is detected by this model.
- Compare to results of default threshold (0.5), the results improve.

Confusion Matrix



Feature Importance - Gradient Boosting Classifier (new threshold 0.5443)

- Most important features:
 1. **Duration** (how long it took the conversation between the sales representative and the potential client)
 2. **Month** (the month of the year)
 3. **Contact** (contact communication type)



Conclusion

Abstract

Data
Understanding

Data
Preparation

Modeling

Results

Marketing Campaign Solution

- Features of potential customer of term deposit account:
 - Customers who talked longer than 7.1 minutes
 - People with balance > \$1,708
 - Students and retired people

