Project 3
Predicting Credit Card Default

Cheng Ji

Content

- The Data
- EDA
- Mine
- Refine
- Model
- Conclusion

The Data

- Credit Card default info.
- Goal: predict the probability of default based on personal and previous payment information.

EDA and Mining

- Clean dataset, no missing values, appropriate data types.
- Rename columns for easier reference.

EDA and Mining

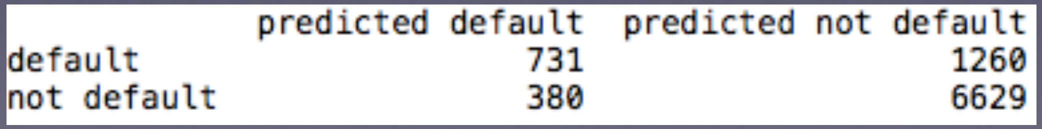
- Plot bill amount variables and payment amount variables. Strong Correlation among those variables, which indicates feature selection.

Refining

- Check correlation between variables to identify potential features
- Standardize numeric features
- Create dummy variables for categorical features
- Train_test_split with stratify

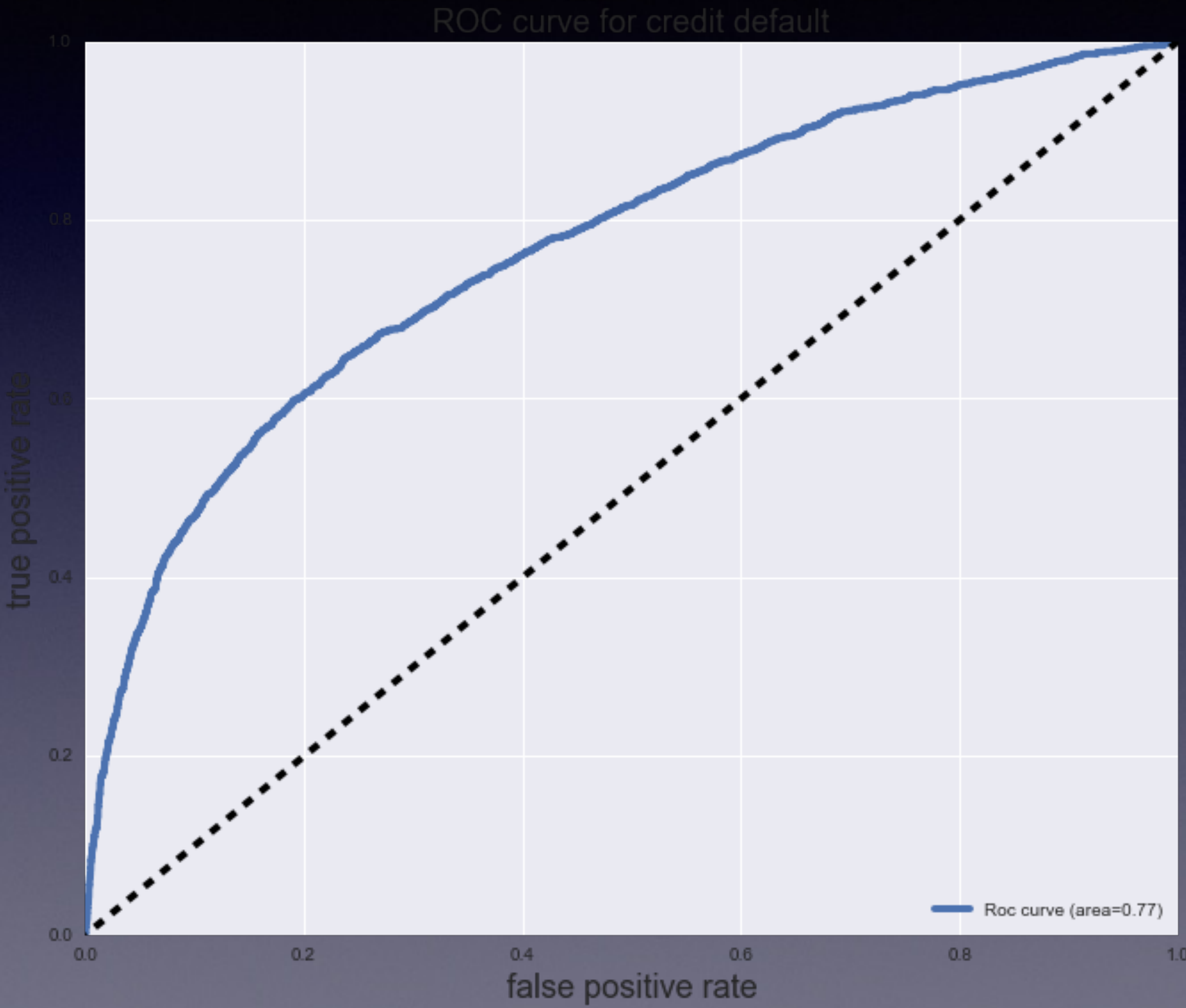
Models

- Logistic Regression with Gradient Descent (SGDClassifier) and Grid Search
- Utilizing Lasso regularization to select features
- F1-score: 0.79, roc-auc: 0.77



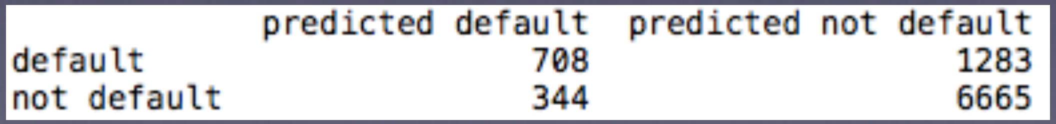
```
             predicted default  predicted not default
default                    731                   1260
not default                380                   6629
```

Models


ROC curve for credit default
1.0
0.8
0.6
0.4
0.2
0.0
true positive rate
0.0
0.2
0.4
0.6
0.8
1.0
false positive rate
Roc curve (area=0.77)

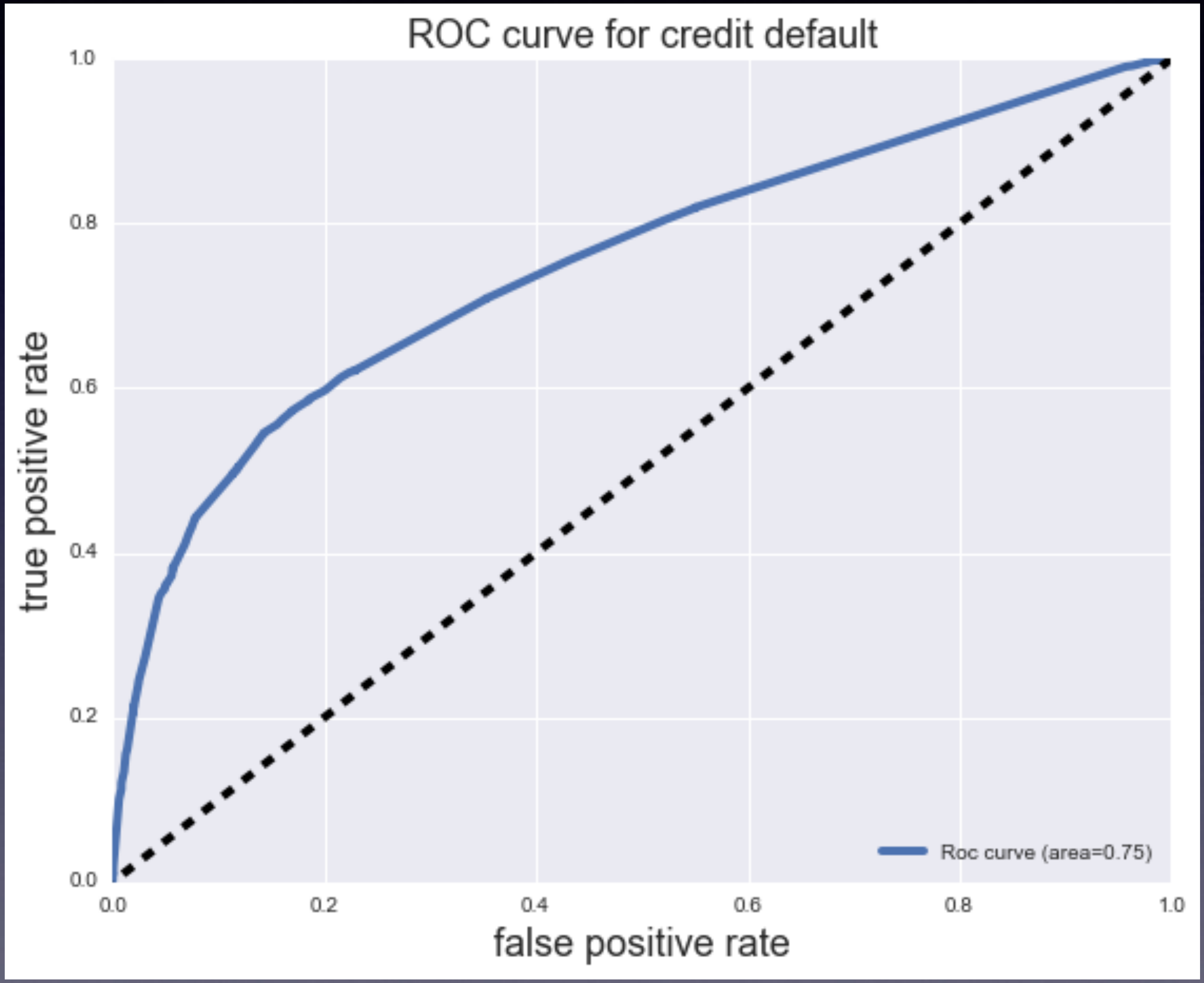
Models

- Logistic Regression with manual feature selection (RFECV) and Grid Search
- F1-score: 0.80, roc-auc: 0.75



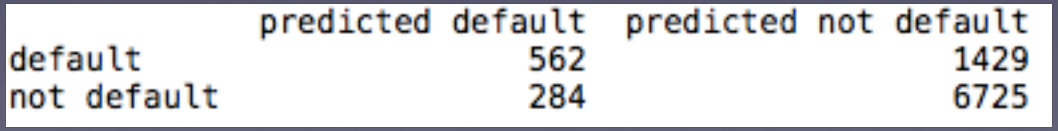
```
             predicted default  predicted not default
default                    708                   1283
not default                344                   6665
```

Models


ROC curve for credit default
1.0
0.8
0.6
0.4
0.2
0.0
true positive rate
0.0
0.2
0.4
0.6
0.8
1.0
false positive rate
Roc curve (area=0.75)

Models

- K Nearest Neighbors with Grid Search
- F1-score: 0.78



```
             predicted default  predicted not default
default                    562                   1429
not default                284                   6725
```

Conclusion

- Potential model skewness due to unbalanced class
- Better dealing with outliers
- Potential multicollinearity
- KNN does not work with feature selection (RFE), and give the worst performance