Home Credit Late Payment Risk Prediction

YUEYING (SHARON) ZHANG

Highlights

Developed an overall understanding of the dataset and the behaviors of consumers with late payment and without late payment through data analysis and visualization.

- Consumers with higher education level are much less likely to have late payment behavior.
- Consumers whose main income source is working are more likely to make a late payment on a loan.

Combined 7 datasets and generated ~100 features.

- Features on individual backgrounds, bureau balance, credit card balance, POS cash balance, installment payment and previous applications.

Trained 3 machine learning models to classify consumers into two groups: with payment difficulty and without payment difficulty

- Random forest, gradient boosting tree, lightGBM.
- LightGBM has the highest test AUC 0.7007.

Review Progress

Completed all stories in Epic1 (explorative data analysis):

- Compared and visualized the features of consumers with late payment and without late payment including education, employment, credits and defaults in their social surroundings.

Completed all stories in Epic2 (classification model training):

- Combined 7 datasets and generated ~100 features
- Used the feature importance functionality of random forest to select 14 features with the highest predictive power.
- Visualized the distributions of selected features.
- Built and tuned 3 binary classification models: random forest, gradient boosting tree and lightGBM considering both model complexity and performance.
- Compared the performance of all classification models with their best hyper-parameter combination and chose the lightGBM model as the final model which has the highest test AUC 0.7007.

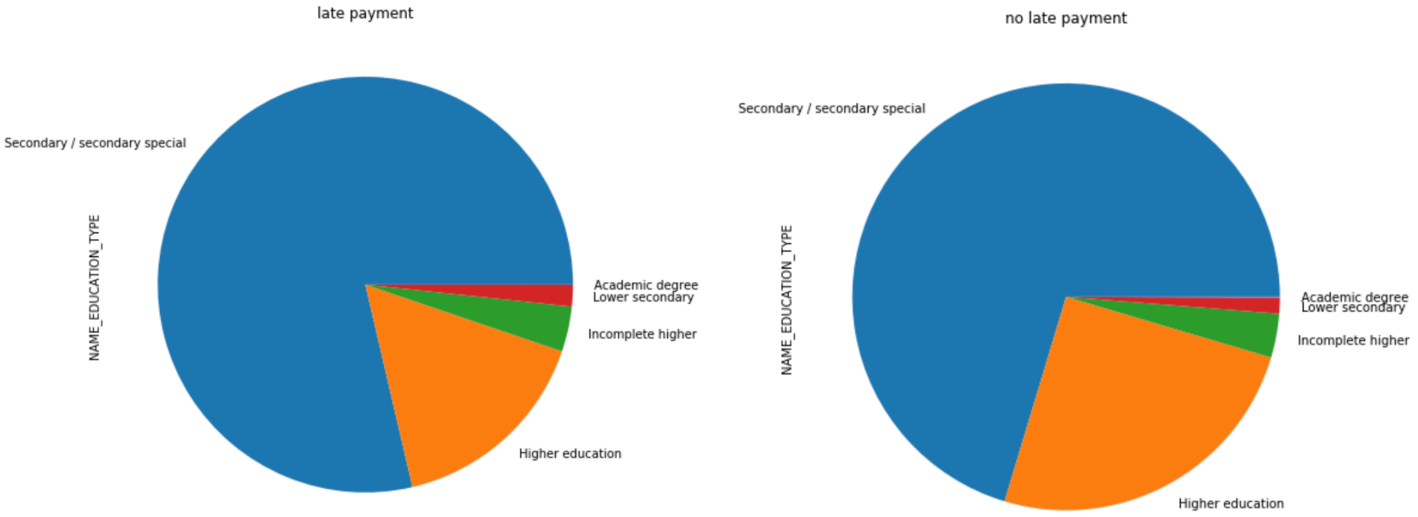
Code Demo and Visualization

Feature Engineering

1.1 Application dataset

```python
def application_feature(df):
    ''' Feature engineer for application dataset
    Args:
        df (dataframe): dataframe of application_train
    Returns:
        df (dataframe): dataframe of application_train with additional features
    '''

    # Remove applications with XNA CODE_GENDER
    df = df[df['CODE_GENDER'] != 'XNA']

    # Categorical features with Binary encode (0 or 1; two categories)
    for bin_feature in ['CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY']:
        df[bin_feature], uniques = pd.factorize(df[bin_feature])
    # One-hot encoding
    df, cat_cols = one_hot_encoder(df, False)

    # Replace DAYS_EMPLOYED: 365243 -> nan
    df['DAYS_EMPLOYED'].replace(365243, np.nan, inplace= True)

    # Engineer new features (percentage)
    df['DAYS_EMPLOYED_PERC'] = df['DAYS_EMPLOYED'] / df['DAYS_BIRTH']
    df['INCOME_CREDIT_PERC'] = df['AMT_INCOME_TOTAL'] / df['AMT_CREDIT']
    df['INCOME_PER_PERSON'] = df['AMT_INCOME_TOTAL'] / df['CNT_FAM_MEMBERS']
    df['ANNUITY_INCOME_PERC'] = df['AMT_ANNUITY'] / df['AMT_INCOME_TOTAL']
    df['PAYMENT_RATE'] = df['AMT_ANNUITY'] / df['AMT_CREDIT']

    return df
```

Consumers with higher education are less likely to make late payment.



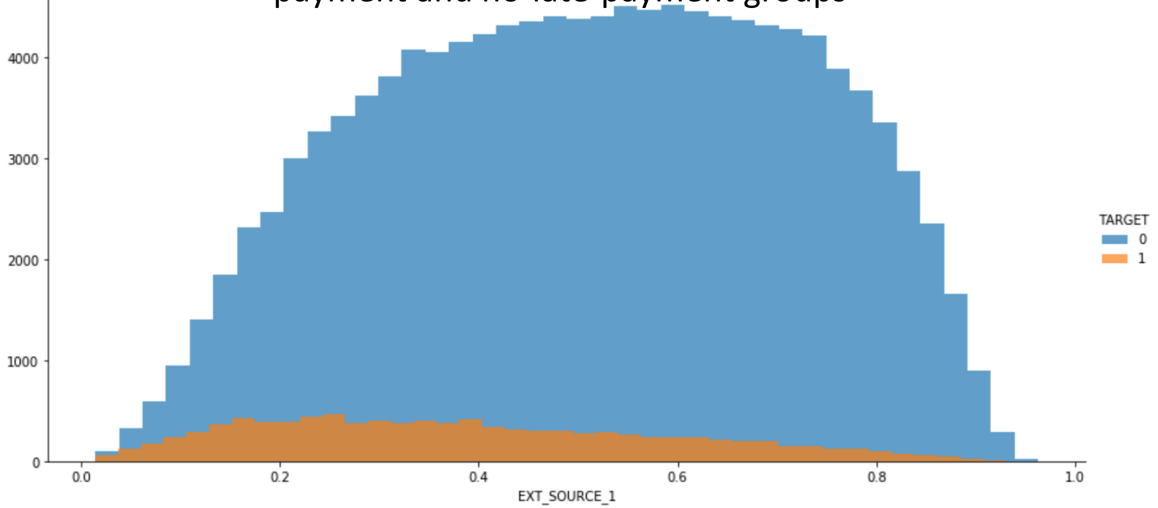
Model Comparison

- Random Forest
- Xgboost
- LightGBM (LightGBM best hyperparamter comes from Kaggle Kernel)

Comparison metric: AUC

```python
def randomForest(X_train, y_train, X_test, y_test):
    """Random forest model training and predicting
    Args:
        X_train (dataframe): dataframe of train set with only selected features
        y_train (dataframe): dataframe of train set with only y label
        X_test (dataframe): dataframe of test set with only selected features used for prediction
        y_test (dataframe): dataframe of test set with only y label to compare with model prediction
    Return:
        rf_fit (a random forest classifier): random forest model trained using X_train and y_train
    """

    # Train random forest model
    rf = RandomForestClassifier(n_estimators=200, max_depth=7, max_features=3)
    rf_fit = rf.fit(X_train, y_train)

    # Predict probability of late payment and print auc on test set
    y_pred_rf = rf_fit.predict_proba(X_test)[:, 1]
    print('auc of random forest: ', roc_auc_score(y_test, y_pred_rf))

    return rf_fit
```

Different distributions of external scores for late-payment and no-late-payment groups



Lessons Learned

Technology

- The data is imbalanced with only 8% of consumers have late payment. As a result, when predicting the label of new consumers, we need to carefully choose the threshold (if the predicted score is above the threshold, we will label the consumer with payment difficulty) instead of using the default 0.5.

Product

- Income is commonly used to evaluate repayment ability by banks. However, there are consumers with high income/credit amount of loan ratio making late payment, and the income/credit ratios of late-payment group and no-late-payment group are very similar. This suggests income may not be a strong factor to identify and predict late payment.
- Consumers who change their identity documents with which they apply for the loan and their phone numbers are more likely to make late payment. This indicates the instability of the customers, which can be strong predictors of late payment behavior.

Recommendations

Need to complete stories in Epic3 (product pipeline, reproducibility and implementation)

- Take user inputs and output prediction scores
- Write unit tests and have all tests passed locally
- Move related data and file to AWS environment
- Write necessary backend structures using Flask
- Design frontend user interface
- Document every file clearly