Summary of FinTech Predictive Model Built

1. Objective

1.1 Problem Statement

Investors often have to choose between hundreds or thousands of available loans at Lending Club. Our objective is to train and evaluate a reliable investment predictive engine to help investors identify the risk of different loans with the 36-month term and screen out the optimal ones to invest in.

1.2 Metrics

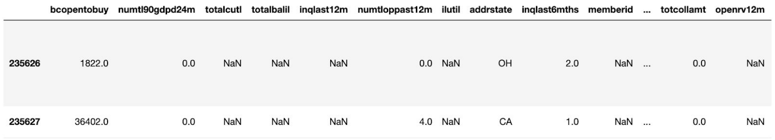
In this project, we will predict whether the status of a loan will be charged-off or default and the probability. It is handled as a binary classification problem. Accuracy is an important metric to evaluate the performance of classification models. However, considering Precision and Recall two aspects together, metrics like AUC or F1 score are more appropriate to be applied to evaluate the performance of our models.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}, \quad F_1 = \frac{2TP}{2TP+FP+FN}, \quad AUC = \frac{\sum_{ins_i \in positiveclass} rank_{ins_i} - \frac{M\times(M+1)}{2}}{M\times N}.$$

2. Data Set

The data of loans launched in 2014 (235,629 rows, 143 columns) will be used for train and test data. The current data of loans listed on 7/13/2017 (103 columns) will be used for prediction. So it is necessary that the current dataset and the historical dataset have the same set of features in the names and the meanings.

Here is the sample data. (More sample data and data dictionary is attached with this assessment.)



	bcopentobuy	numtl90gdpd24m	totalcutl	totalbalil	inqlast12m	numtloppast12m	ilutil	addrstate	inqlast6mths	memberid	...	totcollamt	openrv12m
235626	1822.0	0.0	NaN	NaN	NaN	0.0	NaN	OH	2.0	NaN	...	0.0	NaN
235627	36402.0	0.0	NaN	NaN	NaN	4.0	NaN	CA	1.0	NaN	...	0.0	NaN

3. Data Manipulation and EDA

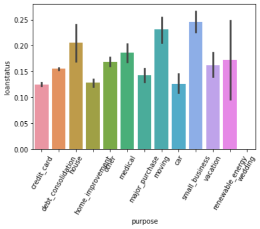
3.1 Data Cleansing

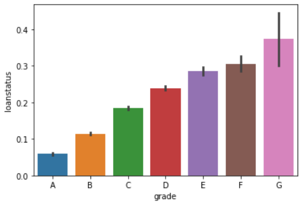
There are 95 common features in both datasets after unifying feature names and meanings. According to our objective of predicting the status of the loans with a 36-month term, we only pick up the data which term = '36 months', loanstatus equals 'Fully Paid' or 'Charged Off'. Because Tree-based models are not sensitive to missing values and outliers, we only fill the missing values with np.NAN.

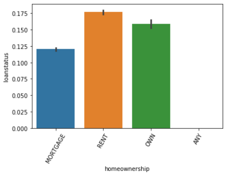
3.2 Feature Selection

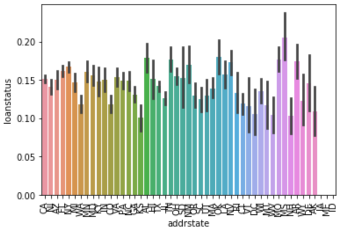
In the 95 columns, there are 19 object features, 76 numeric features. But 18 columns of them are Nulls, and 17 columns have Nulls. Remove the all-null features and the features with only one value, such as term, application_type. Drop duplicated feature fundedamnt (same to loanamnt).

3.3 Data Exploration and visualization









After data exploratory (shown in code), we found nearly 60% of loans are fully paid in 2014 and there are about 15% loans that are 'charged off'. More than 70% of them are used for debt consolidation and credit cards. As grade decreases, the percent of charged-off loans increases. People with mortgages are more likely to borrow money from Lending Club. But for people renting, the percent of charged-off is higher than that of the mortgage class. The distribution of loanstatus varies considerably between states.

3.4 Data Processing

Processing the numerical and categorical features, we get 102 features for modeling.

3.4.1 Numerical Features

After some analysis, we can convert some object features such as intrate, revolutil, loanstatus, emplength to numeric with simple processing (detail shown in code). Feature desc contains complicated content and too many nulls, we convert to 1 for remarks and 0 for np.NAN. Feature earliestcrline records date data, we can compute the time span in month to Dec 2014 as numerical values.

3.4.2 Categorical Features

For categorical features with different characteristics, we apply different methods to encode them, for example, we use label encoding on ordinal features such as grade and subgrade; use frequency encoding on high cardinality features such as zipcode, emptitle and addrstate; use one hot encoding on low cardinality features such as initialliststatus, purpose, verificationstatus and homeownership.

4 Modeling

4.1 Train and Test Dataset

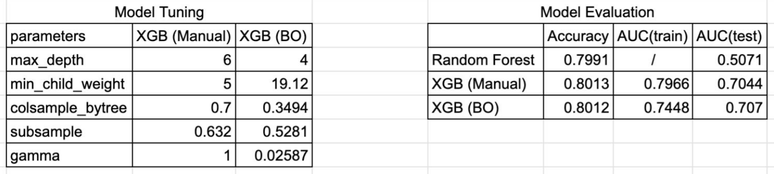
The data of loans issued in the first 9 months will be used as a training dataset to train the model via cross validation. The data in the last 3 months will be used as the testing dataset. The current dataset without labels will only be used to make predictions.

4.2 Model Selection

We choose RandomForest as the benchmark model. It is a highly accurate, robust method and does not suffer from the overfitting problem because of the number of decision trees participating in the process. The mainly predictive algorithm we use is XGBoost, which is a common example of boosting method and outperforms with many advantages in Kaggle, for example, its nature of handling data with heterogeneous features, strong predictive power, robustness to outliers. What’s more, they both have the ability to handle missing values and figure out important features, which helps us to select the most contributing features for the classifier.

4.3 Model Training and Tuning

Without tuning parameters, the accuracy of RandomForest is 0.7991 and the AUC score is 0.5071 which is far from being satisfying. We choose XGBoost models by tuning various parameters manually and automatically with Bayesian Optimization. The best AUC score on the testing dataset is 0.707, which is a great promotion.



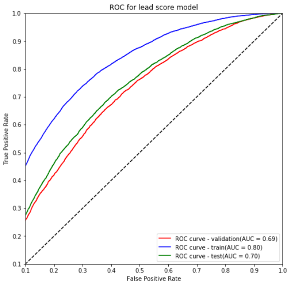
Model Tuning

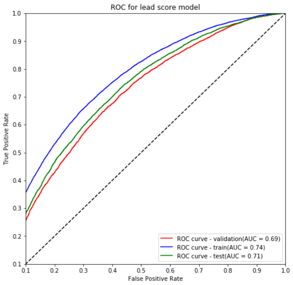
parameters	XGB (Manual)	XGB (BO)
max_depth	6	4
min_child_weight	5	19.12
colsample_bytree	0.7	0.3494
subsample	0.632	0.5281
gamma	1	0.02587

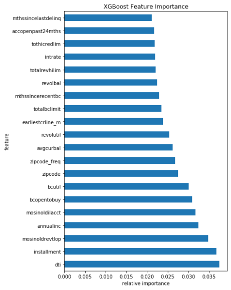
Model Evaluation

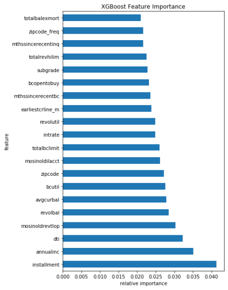
	Accuracy	AUC(train)	AUC(test)
Random Forest	0.7991	/	0.5071
XGB (Manual)	0.8013	0.7966	0.7044
XGB (BO)	0.8012	0.7448	0.707

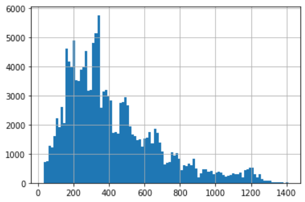
The ROC curves show that xgb(BO) (the right one) is more stable with less variability on the training and testing datasets. The relative important features from xgb(BO) make more sense and are easier to explain. The created features like earliestcrline_m, zipcode_freq are in the top 20.

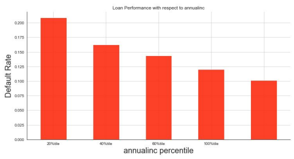


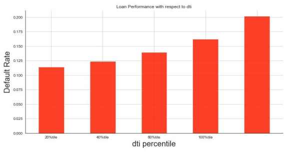


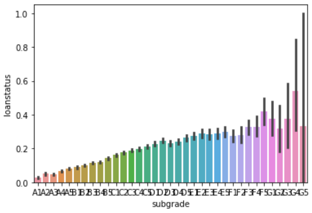












6 Conclusion

After an exploration of data analysis and feature engineering, we trained RandomForest, XGBoost models to predict the status of loans and get a satisfying AUC. Actually, we can do more work in the future to improve it, such as summarize emptitle, analysis desc with NLP methods, create relationship between annualinc and addrstates, handle missing values and outliers more carefully, train the model with more sample data, etc.

7 Appendix (Data Dictionary, Sample Data, Code)