Anjir Team

Credit Default Risk

Predicting how likely each applicant is of repaying a loan?

Goal, identify if a new client shows a high risk for loan default.

How can this help?

Reduce Uncertainty

Proportional Disbursement

Risk Reduction

Doesn't leave business on the table!

Cleaning the data first involves understanding the columns (add infos)

There are 63 clumns with different data types:

- 30 floating point numbers.
- 22 integer numbers – some are numerical categories.
- 11 Strings – Categories involved.

Referral_code has the highest category counts with 7805 categories.

```
preferred_contact          3
referral_code           7805
account_status_code        5
employment_type           16
education                  5
marital_status             3
state                     20
loan_type                 12
loan_purpose               8
origination_channel        4
marketing_campaign        26
dtype: int64
```

Cleaning steps

Since the ML model can't inherently deal with text, the data must be converted to appropriate numbers. Any significant distortion/noise in the model must be removed as much as possible.

- Converting string categorical columns into numerical – Label encoding.
- Converting string categorical columns into numerical and adding new columns to indicate the presence of categorical variables – One hot encoding.
- Replacing illogical outliers with empty values (NAN values).
- Imputing empty cells with the median of the values. In some cases, imputation is approached with a certain grouping.
- Dealing with a few anomalies.
- Changing invalid entries into valid entries.

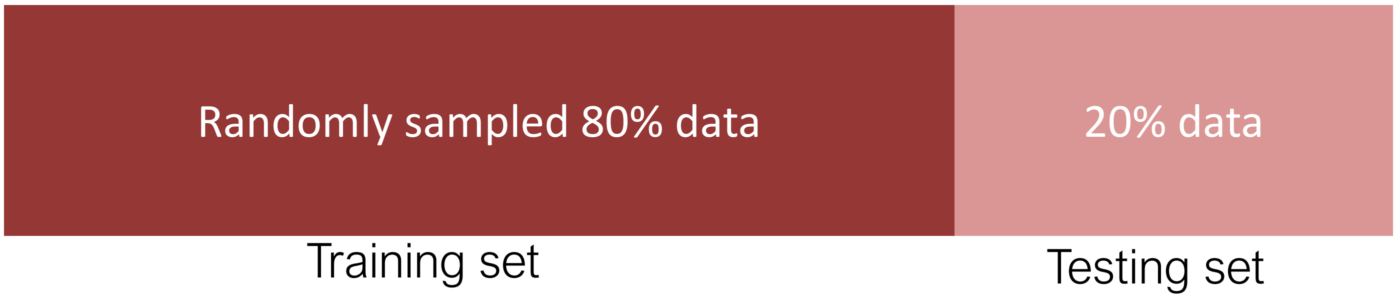
Predictive Modeling – Outcome of the model is expected to identify the potential that someone will default on a loan.

Expected Target Outcome: 0 or 1, 0 – Not a defaulter, 1 – potential defaulter.

Perfomance Metrics used : Accuracy.

Models currently used : Logistic regression, Random forest, XGBoost.

Training and Testing datasets were subjected to the same feature engineering to evaluate the model.



- Out of the main training dataset, a certain percentage is kept untrained to test the model's performance.
- Training set and validation set are split in following percentages: 80% : 20%.
- On the testing set, the target labels are hidden, until the performance is evaluated.

Logictic regression

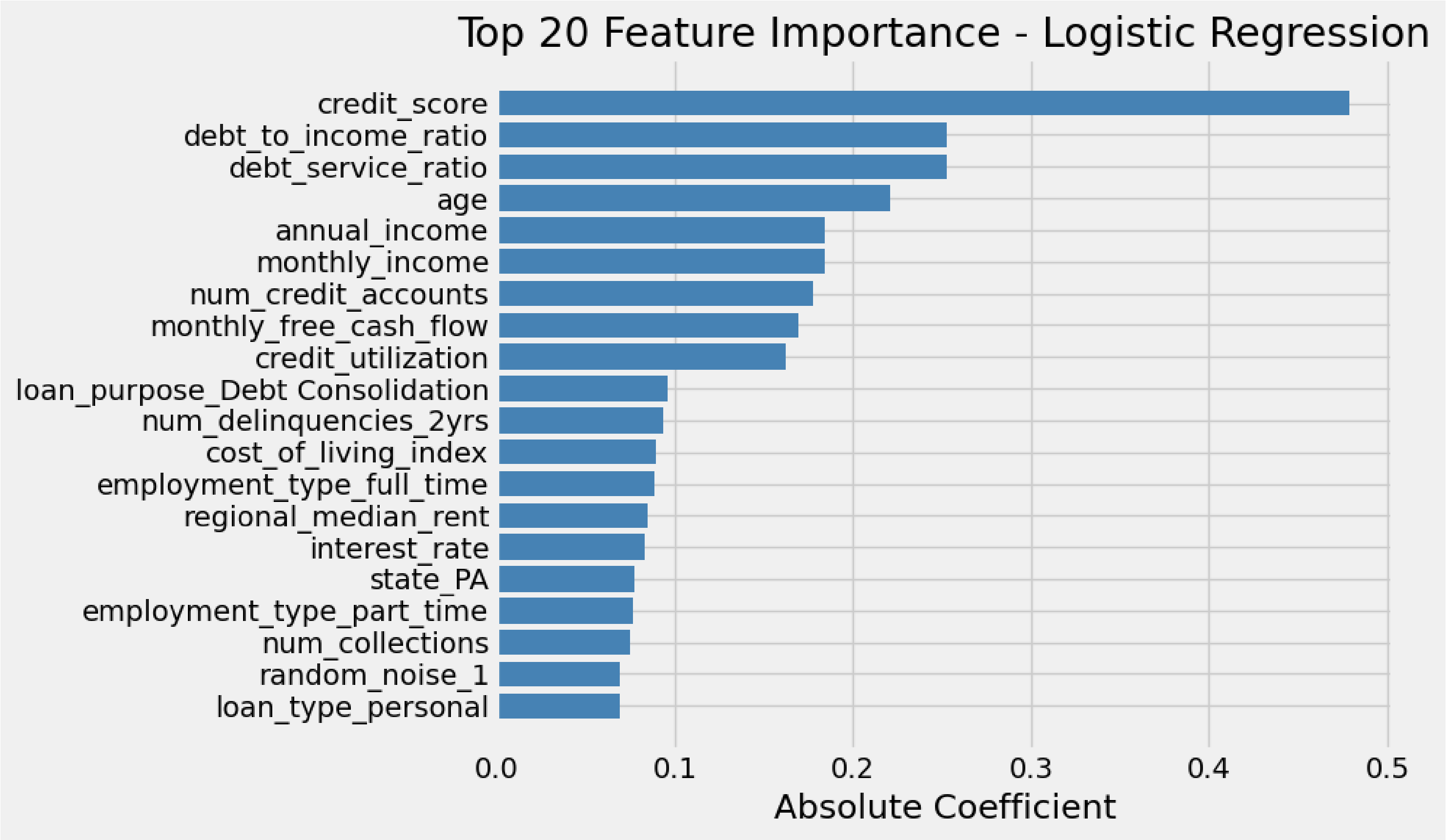
```
The accuracy: 0.7387222222222222
Roc AUC score: 0.7238519289675215
              precision    recall  f1-score   support

           0       0.98      0.74      0.84     17081
           1       0.13      0.71      0.22       919

    accuracy                           0.74     18000
   macro avg       0.55      0.72      0.53     18000
weighted avg       0.94      0.74      0.81     18000

Confusion matrix [[12647  4434]
 [  269   650]]
```

Logictic regression



Random Forest

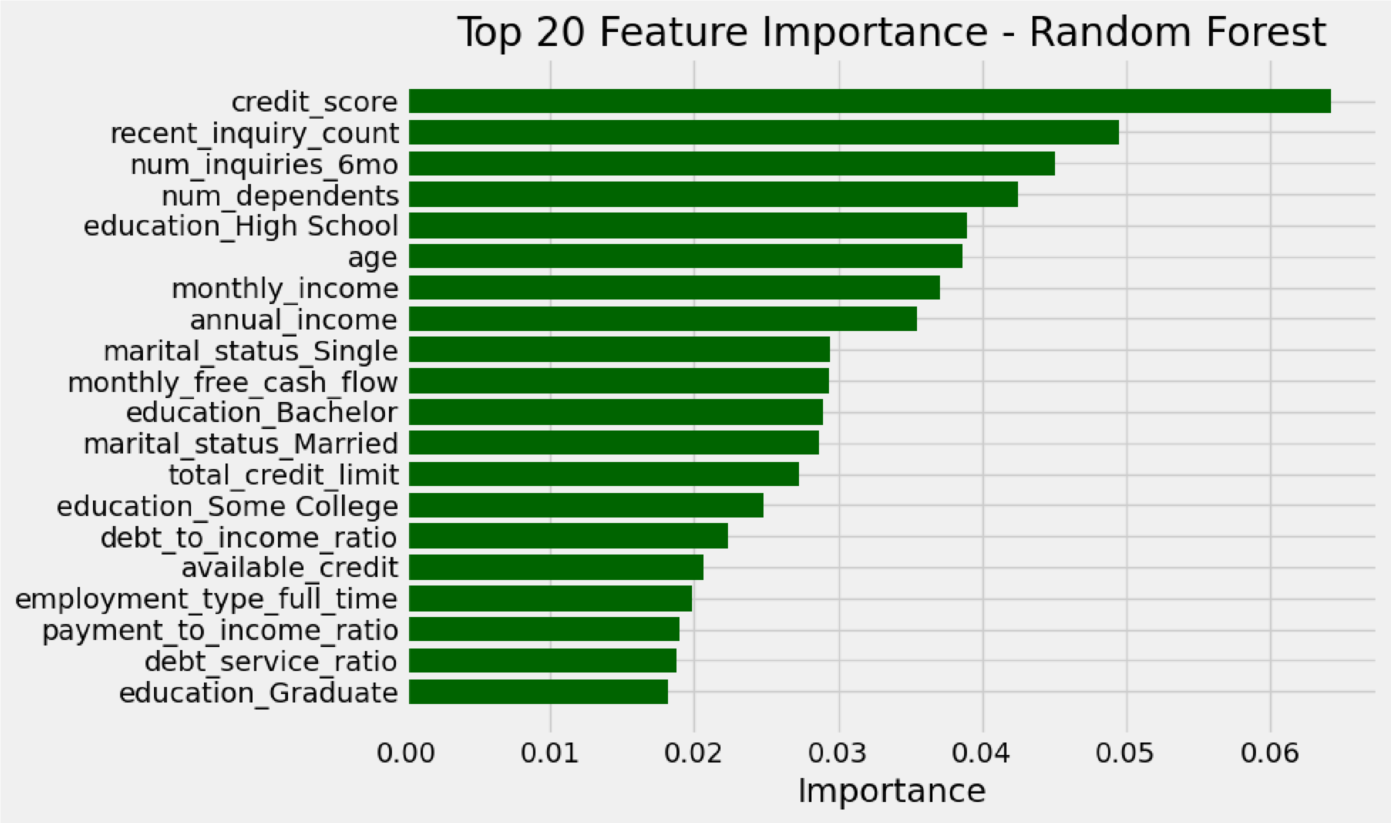
```
The accuracy: 0.9437222222222222
Roc AUC score: 0.5492429370166687
              precision    recall  f1-score   support

           0       0.95      0.99      0.97     17081
           1       0.34      0.11      0.17       919

    accuracy                           0.94     18000
   macro avg       0.65      0.55      0.57     18000
weighted avg       0.92      0.94      0.93     18000

Confusion matrix [[16886   195]
 [  818   101]]
```

Logictic regression



XGBoost

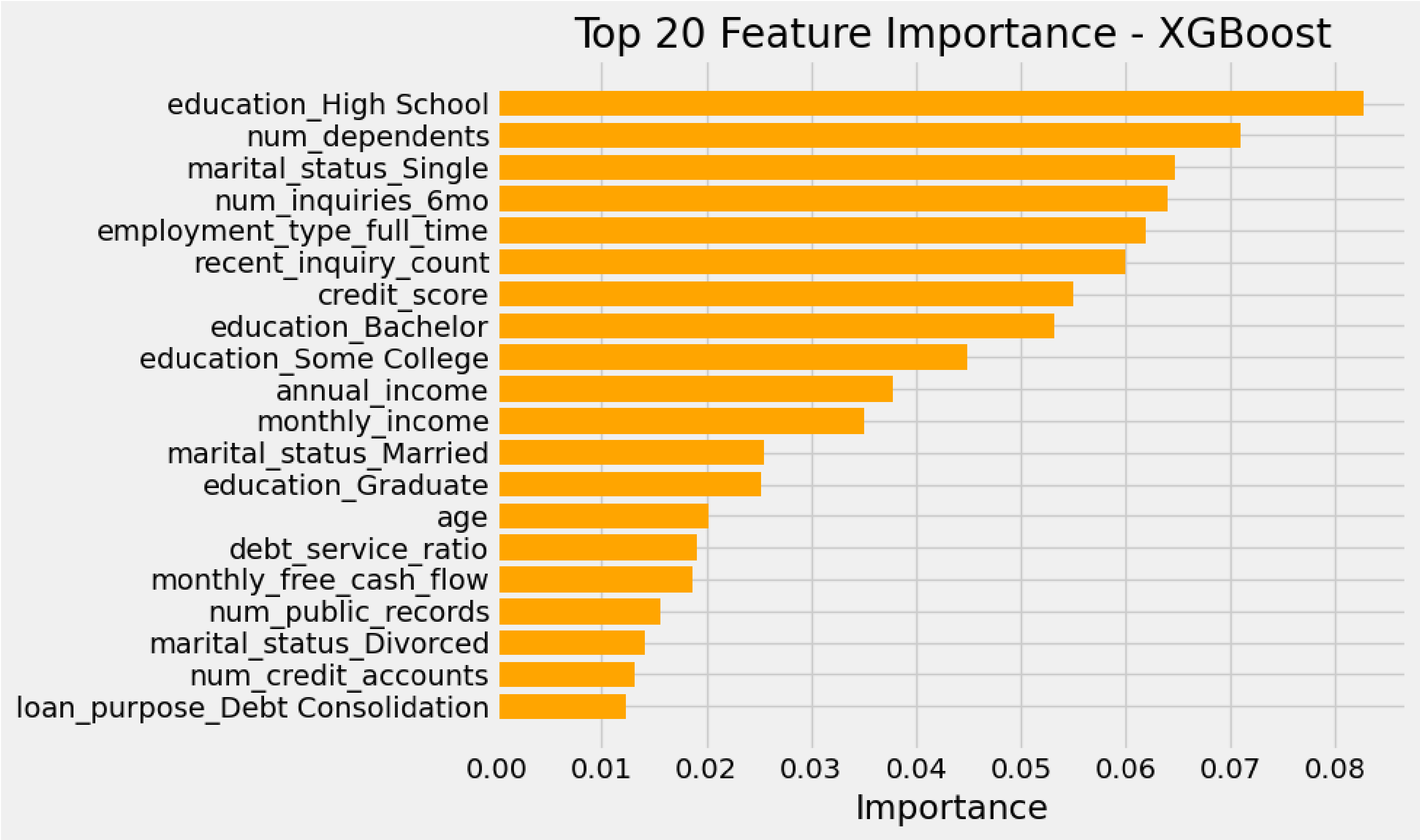
```
The accuracy: 0.8981666666666667
Roc AUC score: 0.665264537737652 6
              precision    recall  f1-score   support

           0       0.97      0.92      0.95     17081
           1       0.22      0.41      0.29       919

    accuracy                           0.90     18000
   macro avg       0.60      0.67      0.62     18000
weighted avg       0.93      0.90      0.91     18000

Confusion matrix [[15794  1287]
 [  546   373]]
```

XGBoost


Top 20 Feature Importance - XGBoost
education_High School
num_dependents
marital_status_Single
num_inquiries_6mo
employment_type_full_time
recent_inquiry_count
credit_score
education_Bachelor
education_Some College
annual_income
monthly_income
marital_status_Married
education_Graduate
age
debt_service_ratio
monthly_free_cash_flow
num_public_records
marital_status_Divorced
num_credit_accounts
loan_purpose_Debt Consolidation
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Importance

Currently there are 3 models used.

Model Name	Hyperparameters	Accuracy
LogisticRegression	(C=2, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1, penalty='l2', random_state=None, solver='liblinear', tol=0.0001, verbose=0, warm_start=False)	74%
RandomForestClassifier	(bootstrap=True, class_weight=None, criterion='gini', max_depth=None, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=500, n_jobs=-1, oob_score=False, random_state=50, verbose=1, warm_start=False)	94%
XGBClassifier	(base_score=0.5, colsample_bylevel=1, colsample_bytree=1, gamma=0, learning_rate=0.1, max_delta_step=0, max_depth=5, min_child_weight=1, missing=None, n_estimators=250, nthread=-1, objective='binary:logistic', reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=0, silent=True, subsample=1)	90%

More Explanation in Notebook.

Random Forest is currently the best chosen model, Until further work.

- More feature engineering could be done, such as mathematical transformation, possibly: certain columns to log and such.
- Advanced techniques like SMOTE could be deployed to handle the class imbalance problems.
- DNN’s can be used.
- Long short term memory networks could be used to incorporate time series data.
- Advanced forms of stacking can be used apart from voting.
- Recursive feature selection can be used to reduce the features.

Questions?

Thank you!