Final Project

Yusen Chen, Lin Cheng, Yihong Guo, Weihang Gao, Yingying Zhuang

Group 28

December 21

Introduction & Data Description

- Presenter: Yusen Chen

Overview

- Introduction of background, motivation and dataset
- Data Preprocessing & Feature engineering
- Methods
- Results & Analysis
- Discussion & Conclusion

Dataset: Credit Card Fraud Detection

- The dataset is to get the insights of Credit Card Defaulters based on the respective features.
  - Data size: 307511
  - Feature number: 122
  - Target: 0, 1.
- Research Goal: Find the best model to predict the decisions of whether to default. Analyze the model performance and the relationship to data.
- This dataset was post by International Institute of Information Technology Bangalore.

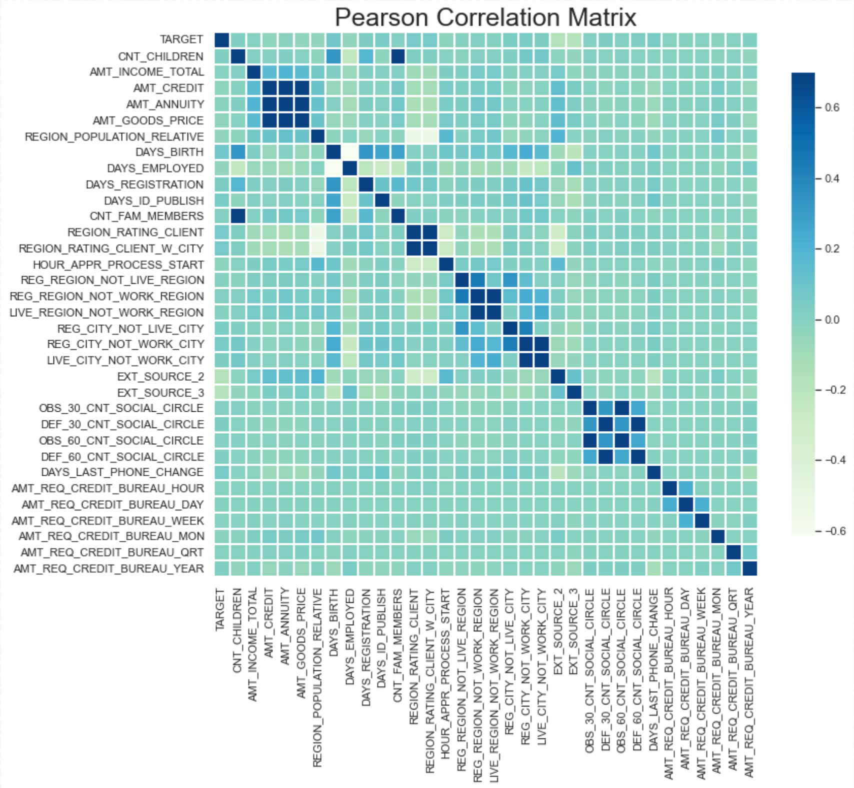
Features' data types

- String variables : Gender, Suite type, Income type, Level of highest education , etc.
- Binary variables : Correctness of application information and Whether a client provided information of certain things and Loan history.
- Integer & Float variables: Amount of loan applied, income, Days of employment, Client's age in days, etc.
- For string variables, use dummy variable for representation.

Distribution of the target

- Imbalanced dataset, employ down-sampling to deal with this problem.

Correlation of Features & Multicollinearity



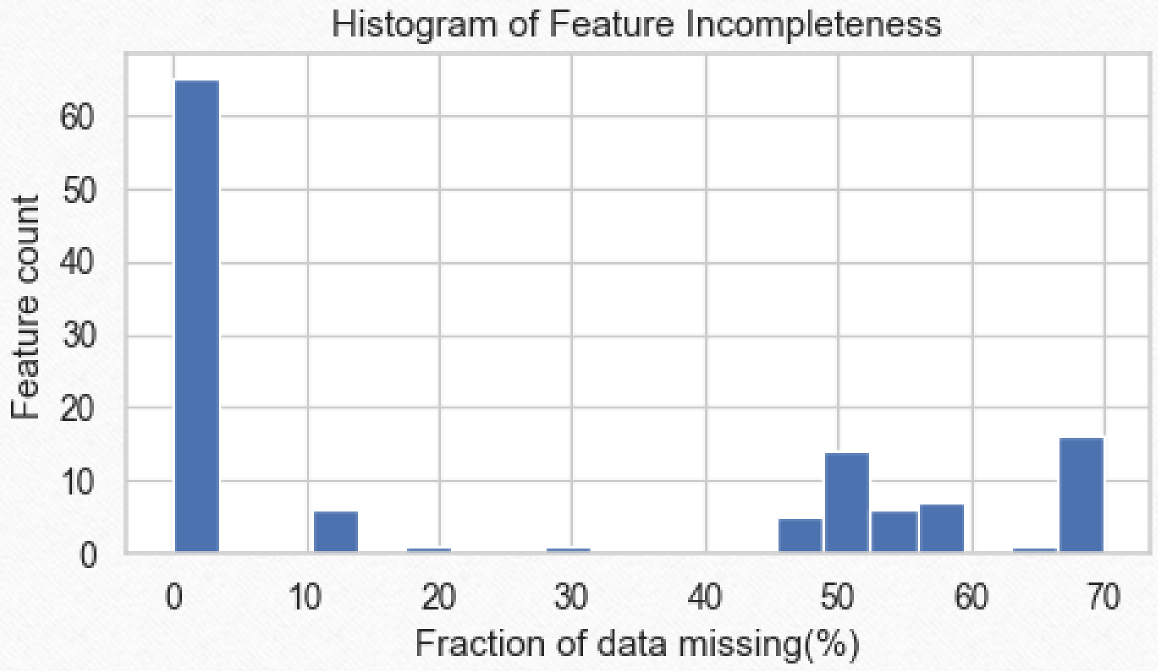
- Calculate variance inflation factor(VIF) value for all variables and drop those with high VIF in linear model

Feature Engineering

- Presenter: Yihong Guo

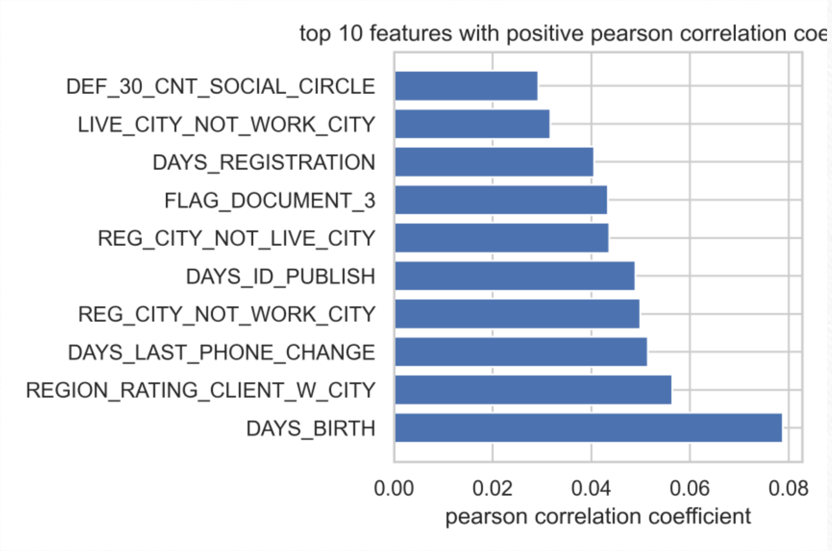
Missing Value Detection and Treatment

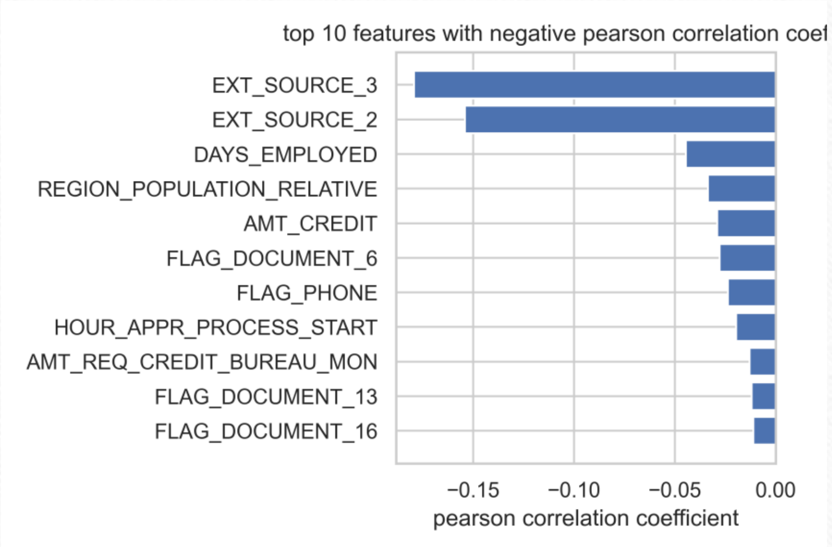
- Drop features with over 30% missing data.
- Fill remain missing data with KNN.



Feature Relationship with Target

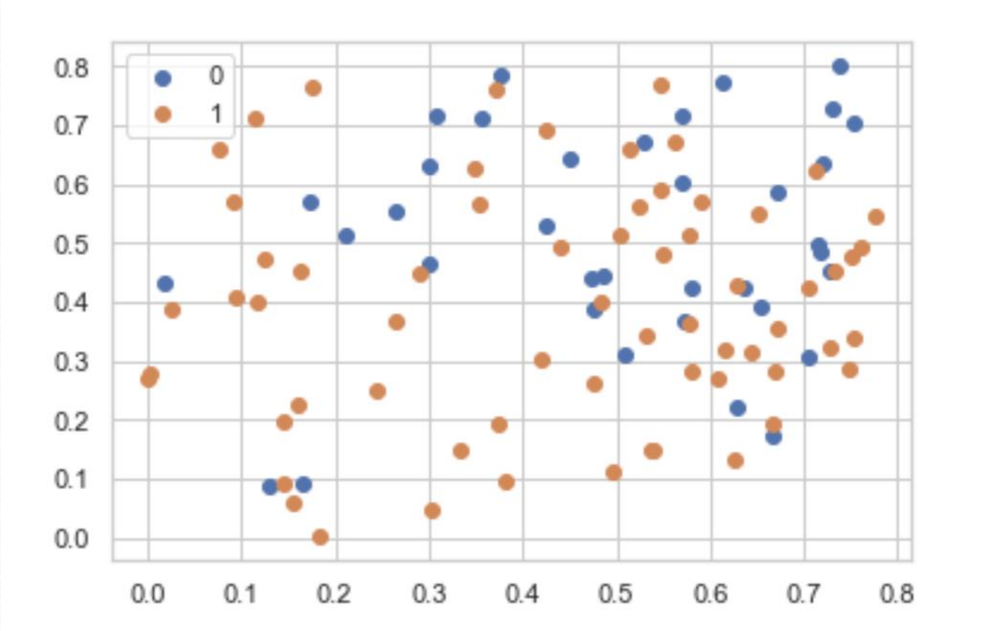
Features are not so corelated to the target.



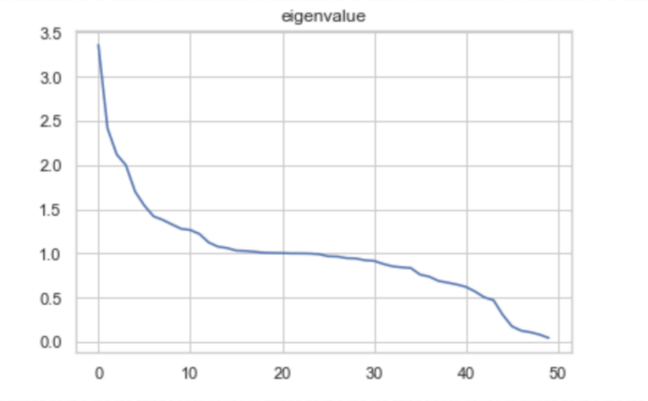


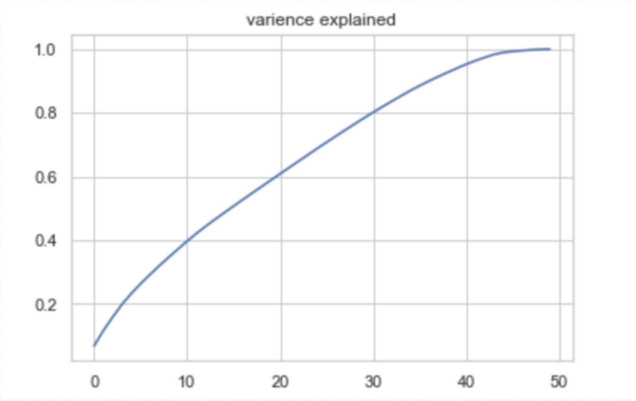
Feature: EXT_SOURCE_2 and EXT_SOURCE_3

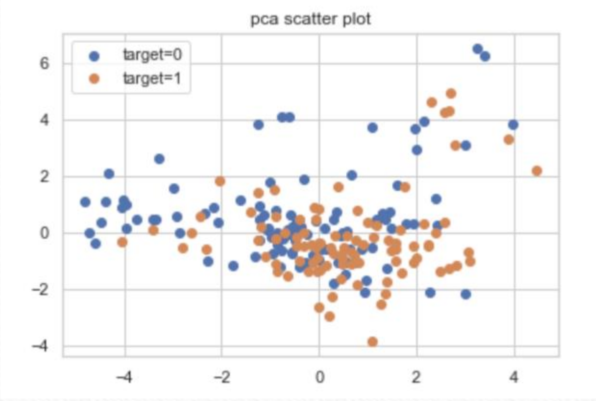
- Scatter plot of the features with target



PCA for Dimension Deduction

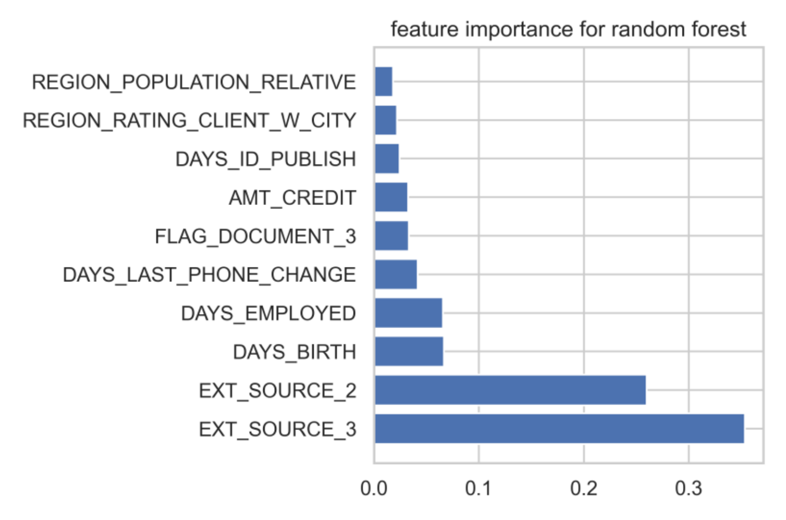


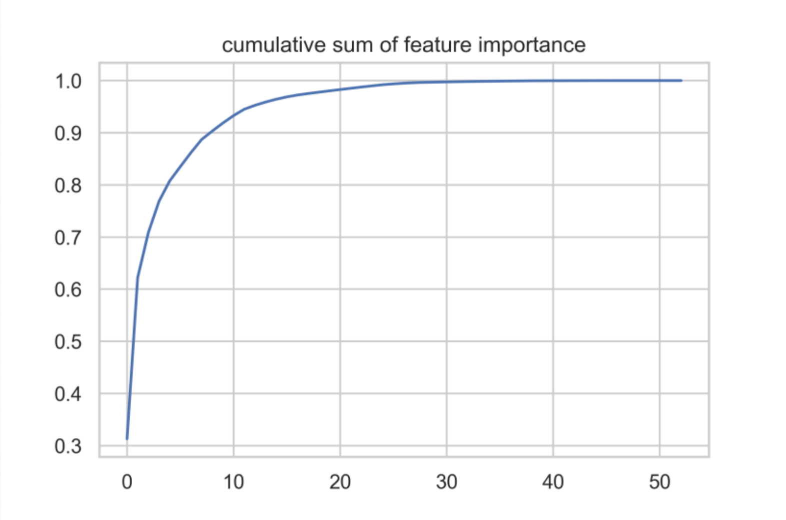




Feature Selection based on Random Forest

- 2 features are more important.
- 10 features contribute over 90% of prediction.





Methods

- Presenter: Weihang Gao

Models

- KNN
- Naïve Bayes, LDA, QDA
- Logistic Regression(with L1 and L2 norm)
- SVM ("Linear" kernel and with "RBF" kernel)
- Tree models
- VGG (CNN)

Evaluation Metrics

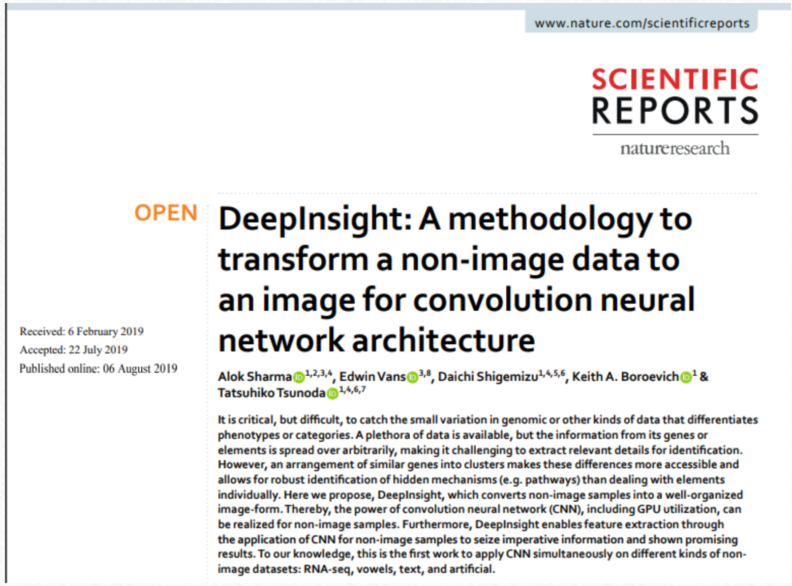
- AUC Score
  - A model with higher AUC is better at predicting True Positives and True Negatives.
- Problem with Accuracy:
  - When dealing with imbalance data, accuracy may go wrong.
- In real practice, banks should focus on precision and recall.

Experiment Settings

- Choose 80% of data as the training set and the rest as testing set.
- Perform down-sampling to the training set.
- Employ Grid Search for parameter tunning.
- Perform experiment with different features selected from Random Forest.
  - 2 important features
  - 10 important features
  - All features
  - 20 principal components

Experiment Settings – CNN

- Problem: We cannot apply CNN on the tabular data directly.
- Transfer tabular data (1D vector) to image data (2D or 3D vector) by **DeepInsight**.
    - DeepInsight is a methodology to transform a non-image data to an image for convolution neural network architecture. The paper is published in 2019.


www.nature.com/scientificreports

SCIENTIFIC REPORTS
natureresearch

OPEN

DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture

Received: 6 February 2019
Accepted: 22 July 2019
Published online: 06 August 2019

Alok Sharma[1,2,3,4], Edwin Vans[3,8], Daichi Shigemizu[1,4,5,6], Keith A. Boroevich[1] & Tatsuhiko Tsunoda[1,4,6,7]

It is critical, but difficult, to catch the small variation in genomic or other kinds of data that differentiates phenotypes or categories. A plethora of data is available, but the information from its genes or elements is spread over arbitrarily, making it challenging to extract relevant details for identification. However, an arrangement of similar genes into clusters makes these differences more accessible and allows for robust identification of hidden mechanisms (e.g. pathways) than dealing with elements individually. Here we propose, DeepInsight, which converts non-image samples into a well-organized image-form. Thereby, the power of convolution neural network (CNN), including GPU utilization, can be realized for non-image samples. Furthermore, DeepInsight enables feature extraction through the application of CNN for non-image samples to seize imperative information and shown promising results. To our knowledge, this is the first work to apply CNN simultaneously on different kinds of non-image datasets: RNA-seq, vowels, text, and artificial.

Experiment Settings – CNN

- Implement VGG-Network by PyTorch.
  - VGG net is a CNN model proposed in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition".
  - Implement VGG with four configurations
    - VGG11, VGG13, VGG16, and VGG 19.
- No feature engineering applied to CNN.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weigh layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-12 conv3-12
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-25 conv3-25 conv3-25 **conv3-25**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-51 conv3-51 conv3-51 **conv3-51**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-51 conv3-51 conv3-51 **conv3-51**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Result & Analysis

- Presenter: Yingying Zhuang

Result for NB, LDA, QDA and KNN

AUC	First 2 features	First 10 features	All features	PCA Result
Naïve Bayes	**0.715**	0.604	0.590	0.504
LDA	0.717	0.724	**0.730**	0.465
QDA	**0.714**	0.712	0.508	0.506
KNN	**0.686**	0.593	0.566	0.449

- Independent assumption for Naïve Bayes is not satisfied as feature number increases.
- LDA is linear supervised dimensional reduction method.
- QDA suffer from collinear as feature number increases.
- PCA lost many information when choosing principal component and it is unsupervised dimensional reduction method.

Result for Logistic Regression

AUC	First 2 features	First 10 features	All features
With L1 norm	0.717	0.724	**0.731**
With L2 norm	**0.717**	0.559	0.560

- L1 normalization can perform feature selection.
- L2 normalization shrink parameters near 0.

Result for SVM

AUC	First 2 features	First 10 features	All features
Linear kernel	0.717	0.724	0.730
RBF kernel	0.703	0.707	0.719

- Linear kernel performs better than RBF kernel.

Result for Random Forest and Xgboost

AUC	First 2 features	First 10 features	All features
Random Forest	0.717	0.729	**0.732**
Xgboost	0.717	**0.728**	0.726

- The two model have similar performances.

Result for CNN

Model	VGG 11	VGG 13	VGG 16	VGG 19
AUC	0.584	0.577	0.594	0.600

- CNN cannot be applied to tabular data as it capture local feature and local features' relation.
- Complicated model leads to overfitting.

Discussion & Conclusion

- Presenter: Lin Cheng

Conclusion

- Random Forests and Logistic Regression have the best performance
  - - 0.73 AUC score.
- Other models have close AUC scores.
  - ~ 0.70 AUC score or even lower.

Model Comparison

- Linear configuration is better than non-linear configuration
  - LDA > QDA
  - SVM Linear Kernel > SVM RBF Kernel
  - Logistic Regression > CNN
- Models with dimensional reduction or feature selection perform well
  - LDA, Logistic Regression with L1 normalization, Tree model

Discussion

- Oversampling vs. Down Sampling
- Overfitting During Cross Validation

Thank you!

Q&A

Individual Contributions

- Yusen Chen: Data cleaning, visualization, feature engineering
- Lin Cheng: Linear Model (Logistic regression with L1 and L2 regularization), parameter tuning, cross validation, down sampling
- Yihong Guo: PCA, feature engineering, Naïve bayes, LDA, QDA, result analysis, model comparison.
- Weihang Gao: Preprocess Data for CNN and implement VGG net
- Yingying Zhuang: Decision Tree Model, Random Forest Tree, KNN Model training and turning.