Credit Card Fraud Detection

Kyle Felkel & Michael Chan

Introduction

- Why is credit card fraud a problem?
  - Significant Financial costs
    - Global costs of $15 billion in 2014
  - Increase in online transactions
    - Losses expected to increase to $30 billion by 2020

Previous Work

- Significant Amount of research has been done
    - Focus is on champion-challenger approach
    - Most papers use Random Forest and Naïve-Bayes models.

Credit Card Fraud: Problem Constraints

- Real time processing required
- The cost of predicting false negatives versus false positives
- Data sets are difficult to find
- Fraudsters adjust, constant retraining required
- Data set imbalance

Our Work

- Examined a real-world dataset
- Selected 2 established Machine Learning algorithms
  - Random Forest, Naïve-Bayes Classifier
- Performed 3 feature selection methods
  - Correlation Filtered, Backwards Selection: Random Forest and Naïve-Bayes
- Compared the results of the 6 models built

Our expectation is that Random Forest and Naïve-Bayes will have similar results. Models built with backwards selected features should also perform better.

Dataset

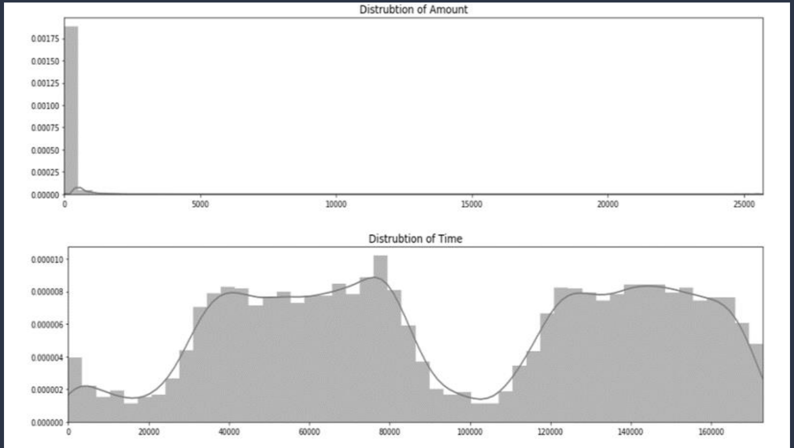
- 284,807 transactions from September 2013 by European cardholders
- 31 Features
  - Amount, Time, v1-v28, Class
  - Features v1-v28 are anonymized and transformed to numeric values for confidentiality

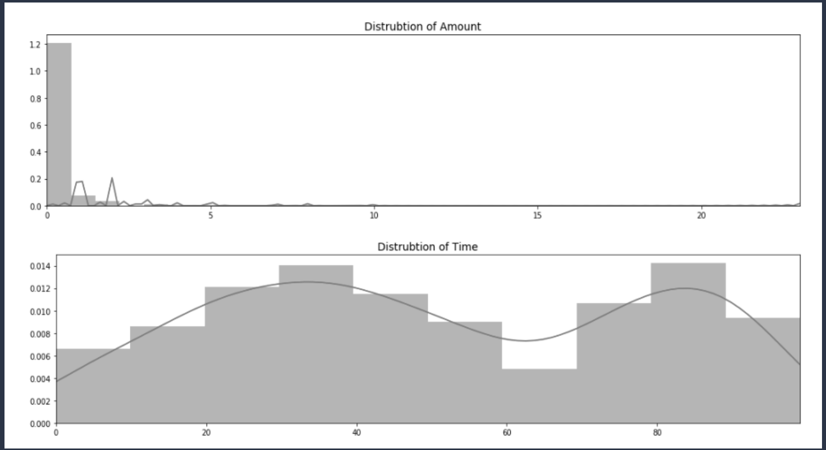
Time	V1	V2	V3	.. V26	V27	V28	Amount	Class
0	-1.35981	-0.07278	2.536347	-0.18911	0.133558	-0.02105	149.62	0
0	1.191857	0.266151	0.16648	0.125895	-0.00898	0.014724	2.69	0
1	-1.35835	-1.34016	1.773209	-0.1391	-0.05535	-0.05975	378.66	0
1	-0.96627	-0.18523	1.792993	-0.22193	0.062723	0.061458	123.5	0
2	-1.15823	0.877737	1.548718	0.502292	0.219422	0.215153	69.99	0
2	-0.42597	0.960523	1.141109	0.105915	0.253844	0.08108	3.67	0

Methodology and Tools

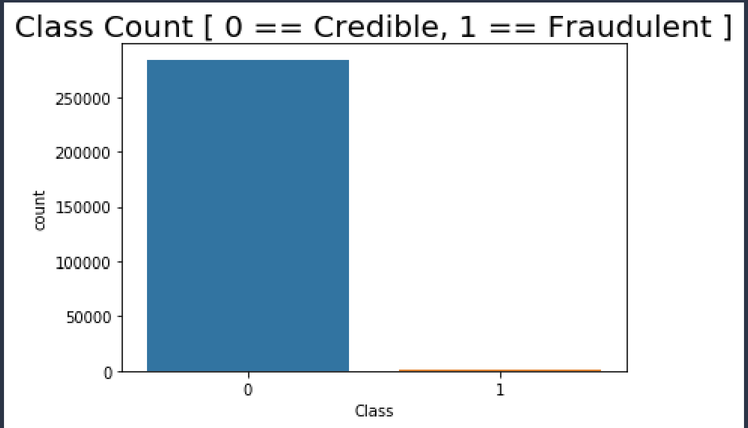
- Data Preprocessing
  - Binning
  - Undersampling
  - Feature Selection
    - Correlation
    - Backwards

Binning: Before and After



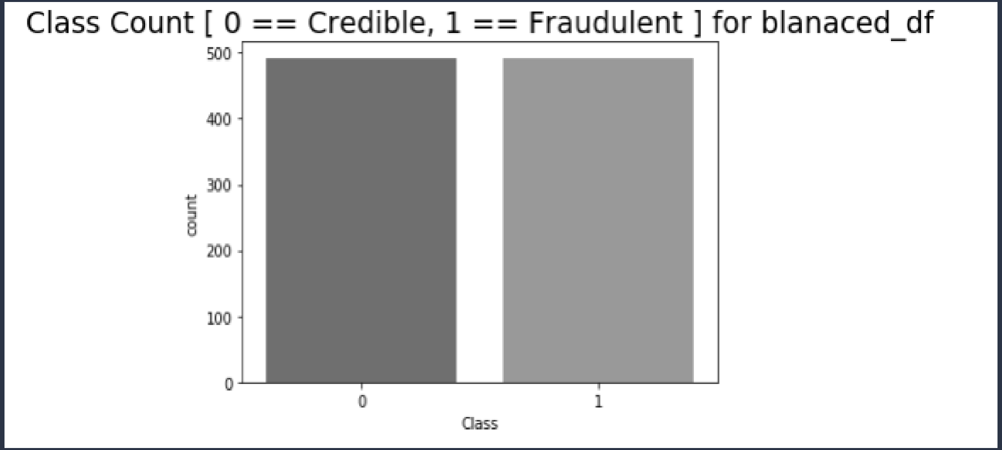


Data Imbalance



- 284,807 Transactions
- 492 Fraudulent transactions
    - 0.17% of all transactions
- Undersampling will be used to correct this

Undersampling



- Creatures a new, relatively small, and balanced data set
    - 984 total Entries
    - 492 Fraudulent Transactions
    - 492 Credible Transactions

Methodology and Tools

- Feature Selection
  - Correlation Filter Feature Selection
    - Features with correlation>0.5 with target feature
  - Backward Feature Selection
    - Build full model and remove features

Methodology and Tools: Machine Learning Algorithms

- Random Forest Algorithm
  - 20 Decision Trees
  - Easy to understand and implement

- Naive-Bayes Algorithm
  - Probability assuming independent features
  - Suitable for real time predictions

Methodology and Tools

- Measures of success
  - False Negatives minimized
  - F1score:
    - Recall = (TP) / (TP + FN)
    - Precision = (TP) / (TP + FP)
    - F1Score = 2 * (Precision * Recall) / (Precision + Recall)
  - Accuracy:
    - Accuracy = (TP + TN) / (TP + TN + FP + FN)

Methodology and Tools

- Tools
  - Python
  - Pandas
  - sklearn/matplotlib/seaborn/mlxtend
  - Jupyter/Anaconda
  - Kaggle

Results:
Feature Selection Outputs

Feature Selection Model:	Selected Features:
Filtered correlation selection	'V3', 'V4', 'V9', 'V10', 'V11', 'V12', 'V14', 'V16', 'V17'
Random Forest Backwards selection	'V4', 'V7', 'V13', 'V14', 'V17', 'V18', 'V21', 'V26', 'V27'
Naïve-Bayes Backwards Selection	'V4', 'V6', 'V7', 'V13', 'V14', 'V19', 'V23', 'V25', 'bin_time'

Results:
Model and Feature Set Combinations

Model No.	Model:	Dataset features:
1	Random Forest	Filtered correlation selected
2	Naïve-Bayes	Filtered correlation selected
3	Random Forest	Random Forest backward selected
4	Naïve-Bayes	Random Forest backward selected
5	Random Forest	Naïve-Bayes backward selected
6	Naïve-Bayes	Naïve-Bayes backward selected

Results: Raw Data

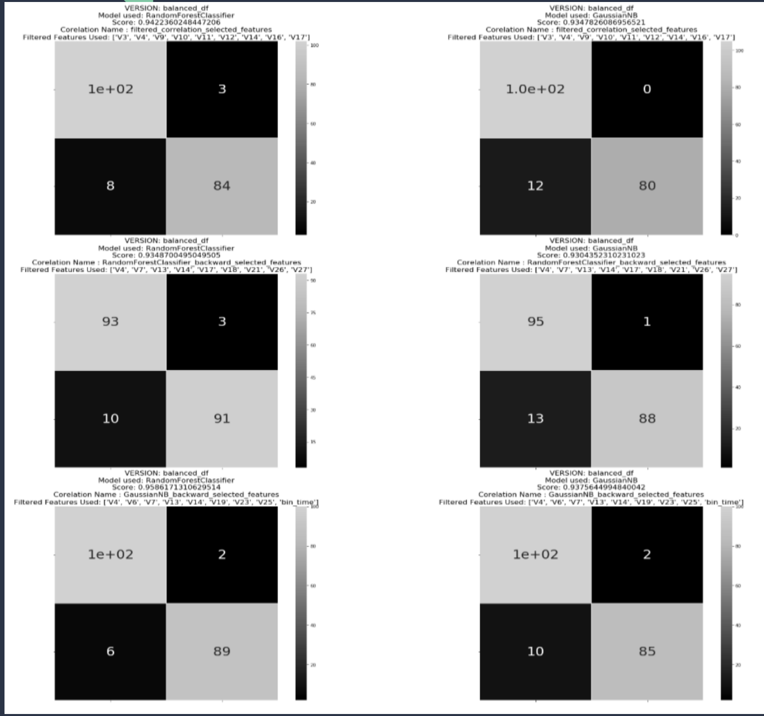
TABLE 3 NUMBER OF CLASSIFICATIONS

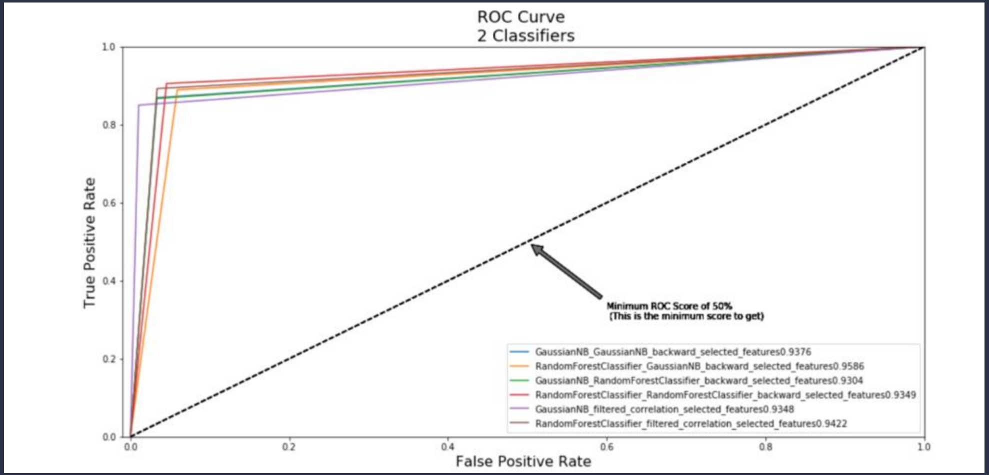
Model	TN	FN	TP	FP
1	84	3	102	8
2	80	0	102	12
3	91	3	93	10
4	88	1	95	13
5	89	2	100	6
6	85	2	100	10

TABLE 4 CLASSIFICATION RATES

Model	TPR	TNR	FPR	FNR
1	.97	.91	.09	.03
2	1.0	.87	.13	0.0
3	.97	.90	.10	.03
4	.99	.87	.13	.01
5	.98	.94	.06	.02
6	.98	.89	.11	.02

Results: Visualization





Results: Scoring

TABLE 5 CLASSIFICATION SCORING

Model	Recall	Precision	F1-score	Accuracy
1	.91	.97	.94	.94
2	.87	1.00	.93	.94
3	.90	.97	.93	.93
4	.87	.99	.93	.93
5	.94	.98	.96	.96
6	.89	.98	.93	.94

Conclusion

- Best model was: Model 5 (Random Forest with Naive-Bayes backwards selected features)
  - FNR: 0.02 F1Score: 0.96 Accuracy 0.96
- All models performed extremely well
- Both Random Forest and Naïve-Bayes suitable algorithms
  - Effect of Algorithm vs Feature Selection

Future work:

- Training/Processing Time
  - Over half a million cards in circulation
- Include Cost-analysis
  - Amount omitted during feature selection
- Thresholding
- More Algorithms

References

[1] S. Patil, V. Nemade and P. Soni, "Predictive Modelling For Credit Card Fraud Detection Using Data Analytics", Procedia Computer Science, vol. 132, pp. 385-395, 2018. Available: 10.1016/j.procs.2018.05.199.

[2] D. Wang, B. Chen and J. Chen, "Credit card fraud detection strategies with consumer incentives", Omega, vol. 88, pp. 179-195, 2019. Available: 10.1016/j.omega.2018.07.001.

[3] N. Carneiro, G. Figueira and M. Costa, "A data mining based system for credit-card fraud detection in e-tail", Decision Support Systems, vol. 95, pp. 91-101, 2017. Available: 10.1016/j.dss.2017.01.002.

[4] F. Carcillo, Y. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection", Information Sciences, 2019. Available: 10.1016/j.ins.2019.05.042.

[5] A. Correa Bahnsen, D. Aouada, A. Stojanovic and B. Ottersten, "Feature engineering strategies for credit card fraud detection", Expert Systems with Applications, vol. 51, pp. 134-142, 2016. Available: 10.1016/j.eswa.2015.12.030.

[6] M. S. Kumar, V. Soundarya, S. Kavitha, E. Keerthika, and E. Aswini, “Credit Card Fraud Detection Using Random Forest Algorithm,” 2019 3rd International Conference on Computing and Communications Technologies (ICCCT), 2019.

[7] M. Zareapoor and P. Shamsolmoali, "Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier", Procedia Computer Science, vol. 48, pp. 679-685, 2015. Available: 10.1016/j.procs.2015.04.201.

[8] J. VanderPlas Python Data Science Handbook : Essential Tools for Working with Data 1st ed. Sebastopol, California: O'Reilly Media, 2016, pp 382-432

[9] A. de Sá, A. Pereira and G. Pappa, "A customized classification algorithm for credit card fraud detection", Engineering Applications of Artificial Intelligence, vol. 72, pp. 21-29, 2018. Available: 10.1016/j.engappai.2018.03.011.

References

[10] S. Carta, G. Fenu, D. R. Recupero, R. Saia "Fraud detection for E-commerce transactions by employing a prudential Multiple Consensus model ", Journal of Information Security and Applications, vol.46, pp.13-22, 2019. Available: 10.1016/j.jisa.2019.02.007

[11] E. Kim et al., "Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning", Expert Systems with Applications, vol. 128, pp. 214-224, 2019. Available: 10.1016/j.eswa.2019.03.042.

[12] J. Han, M. Kamber, & J. Pei Data Mining: Concepts and Techniques, Elsevier Science & Technology 3rd ed. Saint Louis, Morgan Kaufmann, 2011, pp 88-376

[13] P. Cichosz Data Mining Algorithms : Explained Using R 1st ed. Somerset, John Wiley & Sons, Incorporated, 2015, pp 571-587

[14] J. Martinez, "Credit Fraud || Dealing with Imbalanced Datasets," Kaggle, 03-Jul-2019. [Online]. Available: https://www.kaggle.com/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets.

[15] Machine Learning Group of ULB, Credit Card Fraud Detection: Anonymized credit card transactions labeled as fraudulent or genuine, 3, Brussels, Belgium: Kaggle, 2018. [Online] Available: https://www.kaggle.com/mlg-ulb/creditcardfraud

[16] F. Fadaei Noghani and M. Moattar, "Ensemble Classification and Extended Feature Selection for Credit Card Fraud Detection", Journal of Artificial Intelligence and Data Mining, vol. 5, no. 2, pp. 235-243, 2017. Available: 10.22044/jadm.2016.788.

[17] W. Hsu, 'Ch8_Hsu3, California State University Northridge, 2019.