Final Project: Capstone Proposal

Martin L. Gonzalez

Boston University

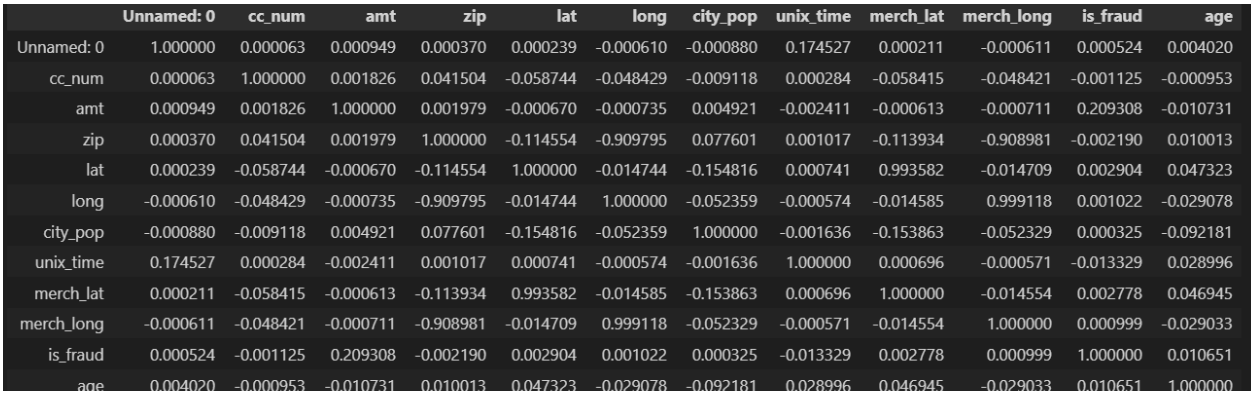
DX699_02 AI for Leaders (summer 1)

Joshua Von Korff

August 3, 2025

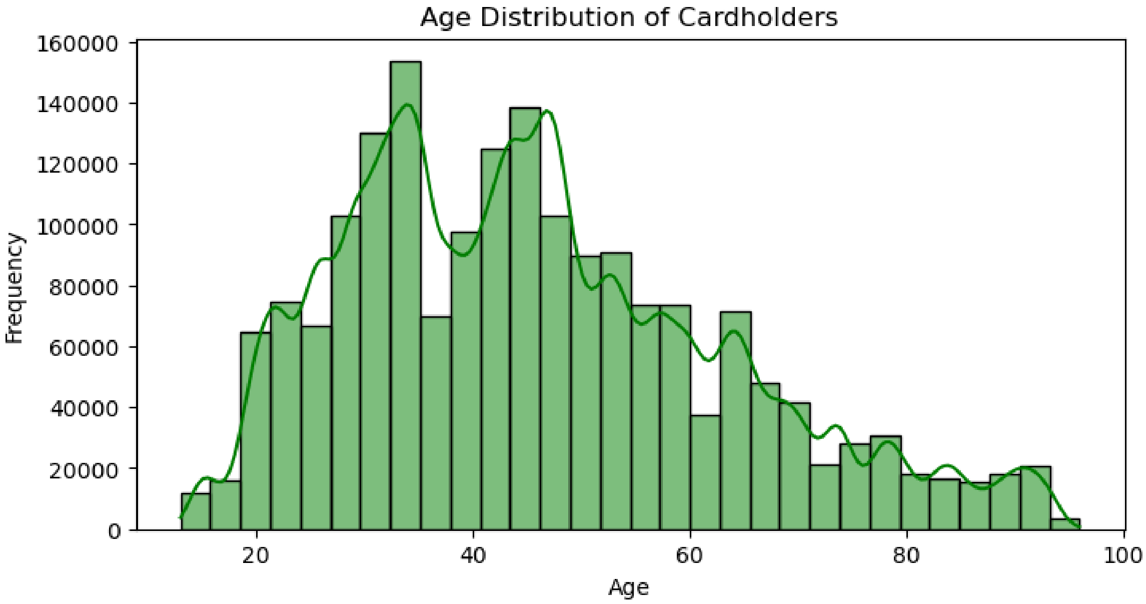
This dataset is based around the Banking, Financial Services, and Insurance sector. The purpose of this dataset is to be able to better understand what makes a transaction fraudulent. Banks usually define a fraudulent transaction by someone other than the account holder or permission from the account owner to make everyday transactions via debit/credit card or through account information. That can be making a purchase in a grocery store via debit or credit card and paying a bill with your account number. Similarly, I had the sole intention of answering "What is fraud?" From this dataset you can answer additional questions such as "What plays a factor into fraud?", "What column has the most weight in determining a fraudulent transaction?", "What is a Fraudulent transaction?", and "How can we prevent a fraudulent transaction?" Fraud is an issue almost every bank and insurance company face, whether that be credit fraud, transactional fraud like what this dataset focuses, or insurance fraud - which can include car, accident, health, home, auto, and more. For this dataset we will focus on transactional fraud usually seen in banks. The data is from kaggle and it is a real world representation of credit card transactions. There are 1852,394 rows, 23 features (columns), and the target variable is binary is_fraud=1 for fraud and is_fraud=0 for a real payment (Bhadouria, n.d.)

Weeks 1-7 were weeks dedicated to more exploratory data analysis while weeks 8-12 were focused more on the models itself. We had lessons on regression and tree based models such as linear regression, random forest, KMeans, and more. The univariate and bivariate analysis from weeks 1-7 helped tremendously with this project. Being able to measure each feature at a time or several at a time helped in understanding the relationships between the variables. The use of heatmaps helped to find the correlation between the features and the target variable.



	Unnamed: 0	cc_num	amt	zip	lat	long	city_pop	unix_time	merch_lat	merch_long	is_fraud	age
Unnamed: 0	1.000000	0.000063	0.000949	0.000370	0.000239	-0.000610	-0.000880	0.174527	0.000211	-0.000611	0.000524	0.004020
cc_num	0.000063	1.000000	0.001826	0.041504	-0.058744	-0.048429	-0.009118	0.000284	-0.058415	-0.048421	-0.001125	-0.000953
amt	0.000949	0.001826	1.000000	0.001979	-0.000670	-0.000735	0.004921	-0.002411	-0.000613	-0.000711	0.209308	-0.010731
zip	0.000370	0.041504	0.001979	1.000000	-0.114554	-0.909795	0.077601	0.001017	-0.113934	-0.908981	-0.002190	0.010013
lat	0.000239	-0.058744	-0.000670	-0.114554	1.000000	-0.014744	-0.154816	0.000741	0.993582	-0.014709	0.002904	0.047323
long	-0.000610	-0.048429	-0.000735	-0.909795	-0.014744	1.000000	-0.052359	-0.000574	-0.014585	0.999118	0.001022	-0.029078
city_pop	-0.000880	-0.009118	0.004921	0.077601	-0.154816	-0.052359	1.000000	-0.001636	-0.153863	-0.052329	0.000325	-0.092181
unix_time	0.174527	0.000284	-0.002411	0.001017	0.000741	-0.000574	-0.001636	1.000000	0.000696	-0.000571	-0.013329	0.028996
merch_lat	0.000211	-0.058415	-0.000613	-0.113934	0.993582	-0.014585	-0.153863	0.000696	1.000000	-0.014554	0.002778	0.046945
merch_long	-0.000611	-0.048421	-0.000711	-0.908981	-0.014709	0.999118	-0.052329	-0.000571	-0.014554	1.000000	0.000999	-0.029033
is_fraud	0.000524	-0.001125	0.209308	-0.002190	0.002904	0.001022	0.000325	-0.013329	0.002778	0.000999	1.000000	0.010651
age	0.004020	-0.000953	-0.010731	0.010013	0.047323	-0.029078	-0.092181	0.028996	0.046945	-0.029033	0.010651	1.000000

From this correlation matrix we are able to see what numeric features are most relevant to use. From this, the top feature correlated with fraud is the “amt” column, which is the amount of the transaction. What is shocking is this shows us that features like “age” are very weak which tells us that it does not matter how old you are to become a victim of fraud. I was surprised to see most of the features are around the 0 mark, meaning there is no correlation to the target variable. I would have assumed the location of where this fraudulent transaction has taken place played a huge factor. From there we use the following chart to tell us more about the “age” column.



This chart specifically shows us the age distribution of individuals who hold a debit or credit card. Where the frequency being the amount of individuals in this column.

For this project I also utilized the “RandomForestClassifer” model. It's a collection of decision trees which will help improve model accuracy and with generalization. The importance of this model derives from its feature selection. Not only does it use different subsets of the data per tree, it will also introduce randomness when choosing those variables, creating a sense of fairness in the model (Jain, 2025).

With this dataset we can answer the questions: Which features are most predictive of fraud? How does the fraudulent transaction differ from merchants, locations, or times of day? Are there certain amounts or are there any patterns that are more likely to be fraudulent? Due to the dataset having features such as “trans_date_trans_time” which has the day, month, year, hour, and minute - we are able to get an accurate measure on what is most common. Could it be during holidays? Or summer months where everyone is vacationing. My prediction is it is more common around the summer months, when people go away on vacation especially in another country. That is more prone to pick pockets and fraudsters. Additional questions can include: What model is best suited for this data? What model is most accurate? And what performance metrics can we use for this dataset. Ideally, for a supervised technique you can use models such as logistic regression, decision trees or random forests, XGBoost, and neural networks. For an unsupervised learning you can use models such as autoencoders, isolation forests, and DBSCAN for anomaly detection (Bhadouria, n.d.).

For Module C the models that would be best suited for this dataset are binary classification related models. Models such as random forest, decision trees, logistic regression, etc. For the next session I hope to run these models and other binary classification models such as Gradient Boosting models along with cluster based models such as K-Nearest Neighbors to get a similarity based prediction. The model I am excited to run would be a neural network model, much like the brain's thinking process. I like to visualize how the model came up with its answer. I will measure the success of these models by the performance metrics such as accuracy, F1 score, the AUC-ROC curves, as well as precision and recall. I will also visualize the confusion matrix table so it is easier to see the true positives (TP), true negatives (TN), False Positives (FP), false negatives (FN). I will utilize feature selection techniques we have learned in this course to fine tune the model and rule out the less important features which will enhance the models performance. The end goal with Module C would be to answer the same questions we have had in this course. By the end of Module C I would have run more advanced data science models that would show more detail in both the metrics and visualizations. With the clustering algorithms I will be able to determine any similarities in the data. I predict a lot of the locations will be grouped together - locations such as similar zip codes or neighborhoods closely put together.

I hope to build off on this topic into a potential thesis whether that be for a masters or PhD program. In the end these questions will help the financial services industry get a more in depth understanding of fraudulent transactions. People will then be able to come up with safety measures and protocols to prevent fraud from happening, not just for banks but for insurance and other financial sectors.

My groupmates and I worked greatly together. We maintained communication throughout the semester even though we worked on our own datasets the times we did talk, we were efficient in what we needed to say and everyone responded respectfully and quickly. The feedback we received when we had to submit weeks 2, 4, and 9 were very useful. It gave great guidance into what I need to adjust for my remaining homework assignments as well as the final project.Team members Michael Cockrell, Jazmiin Deleon, Cory Fournier, Stephen Kozar, and I Martin Gonzalez will be ranked a five. The group members were great with their response time and their feedback. A group member accidentally submitted the wrong file for one of the feedback fruits, that group member realized their mistake very quickly and soon emailed the group owning up to their mistake and also attached the correct submission.

In conclusion, the goal of this dataset is to predict whether a transaction is fraud or not. We can do that by various binary classification models such as logistic regression, which predicts 0/1, true/false, yes/no output. Weeks 1-7 revolved around explanatory data analysis which helped in determining the relationships between the features of the dataset. While weeks 8-12 were model focused, teaching us different types of models and how each model specifically works. For the next class I will run more classification and clustering algorithms to further determine more information with this dataset. I will utilize more features as well, with an emphasis on the date/time feature to answer the question “Does time of year matter with fraud?”. Lastly, the feedback fruit was a great addition to this class as it gave another set of eyes onto what I can do better with my code and answers to different types of questions.

References

Bhadouria, T. (n.d.). *Credit Card Fraud Detection*. Kaggle. https://www.kaggle.com/datasets/tusharbhadouria/credit-card-fraud-detection

Jain, S. (2025, July 23). *Interpreting Random Forest Classification Results*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/interpreting-random-forest-classification-results/