STA303 PROJECT PROPOSAL

Danish Ahmed Bombal

Student ID: 1007733008

RESEARCH QUESTION

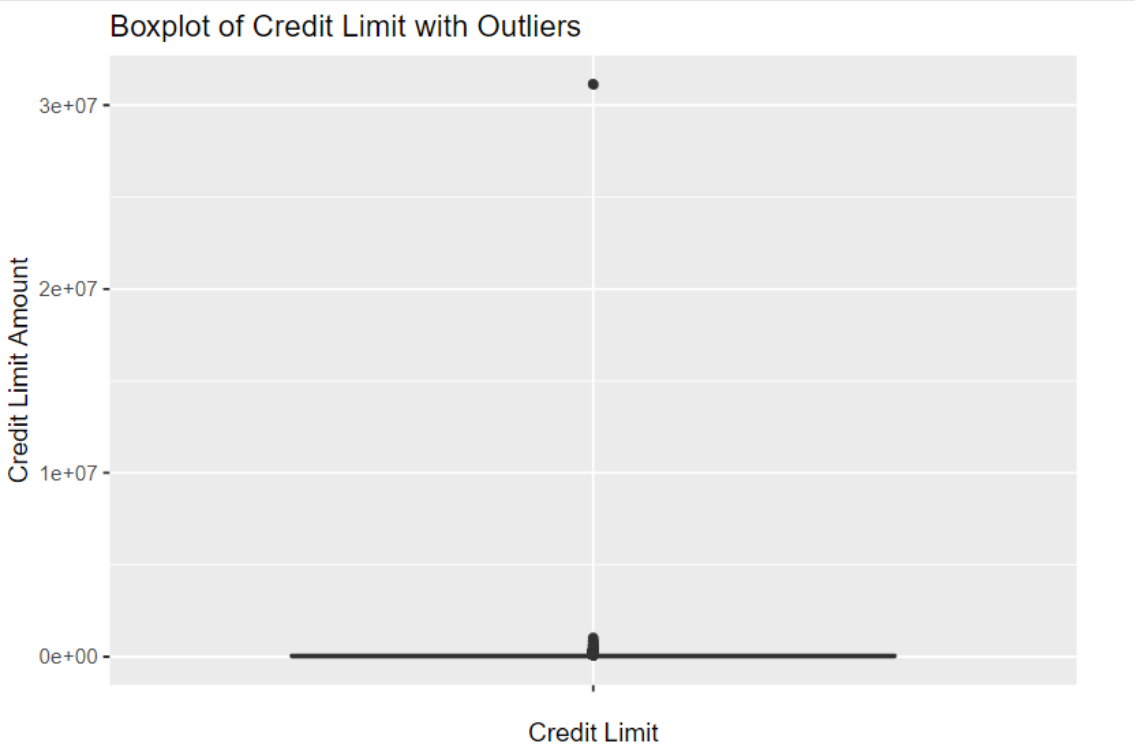
- What are the key predictors of credit card default among American Express customers, and how does the interaction between demographic factors impact the likelihood of default?
- Double Major in Statistics and Economics with a focus in Data Analytics alongside a minor in Mathematics.
- Passionate about the financial industry particularly the banking sector.
- Credit default prediction is a central risk mitigation technique in consumer lending business.
- Allows for optimize lending decision leading to a better consumer experience.

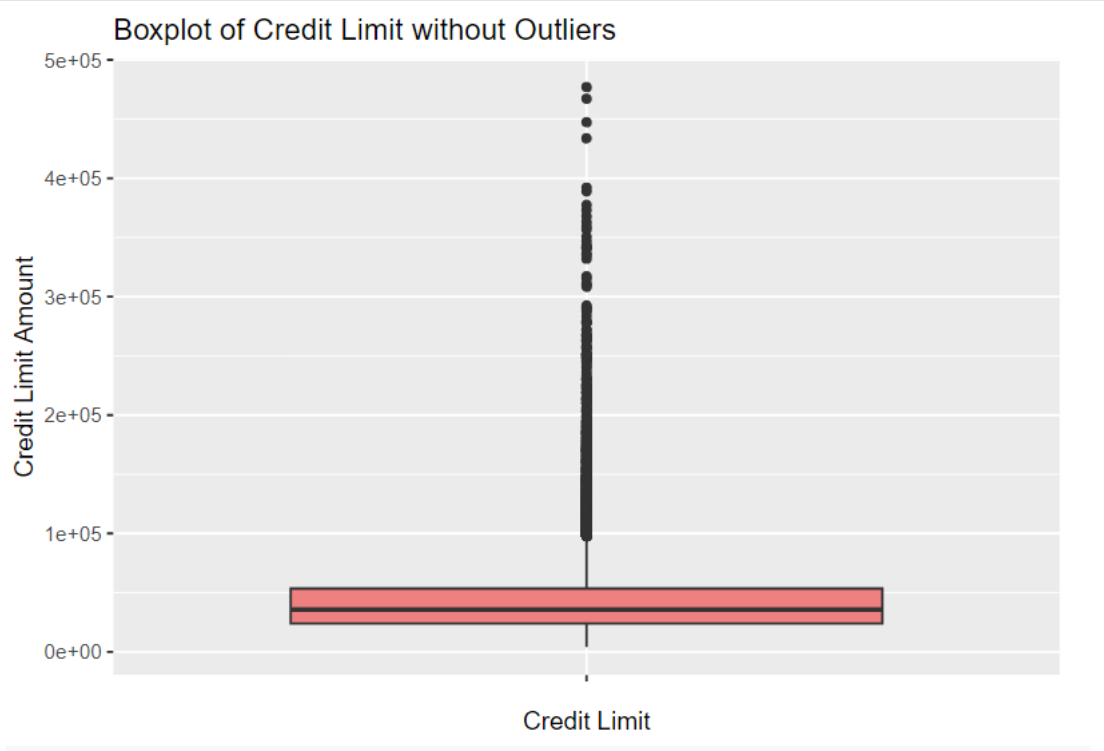
LITERATURE REVIEW

- Paper 1: Zhou, X., Zhang, W., & Jiang, Y. (2020). Personal credit default prediction model based on convolution neural network. Mathematical Problems in Engineering, 2020, 1–10. https://doi.org/10.1155/2020/5608392
- Paper 2: Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: a machine learning methodology. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-98361-6
- Utilized data of peer-to-peer lending business platforms in Taiwan. Used predictors like income, job, marital status and residence
- Machine learning models like gradient boosting model, support vector machines, random forest and convolution neural networks.
- Paper 3: Sayjadah, Y., Hashem, M., Alotaibi, F., & Kasmiran, K. A. (2018). Credit Card Default Prediction using Machine Learning Techniques. IEEE Xplore. https://doi.org/10.1109/icaccaf.2018.8776802
- Paper 4: Calabrese, R., & Osmetti, S. A. (2011). Generalized Extreme Value Regression for Binary Rare Events Data: an Application to Credit Defaults. RePEc: Research Papers in Economics. https://econpapers.repec.org/RePEc:ucd:wpaper:201120
- Paper 5: Kealhofer, S. (2003). Quantifying Credit Risk I: Default prediction. Financial Analysts Journal, 59(1), 30–44. https://doi.org/10.2469/faj.v59.n1.2501

EXPLORATORY DATA ANALYSIS: HIGHLIGHTS

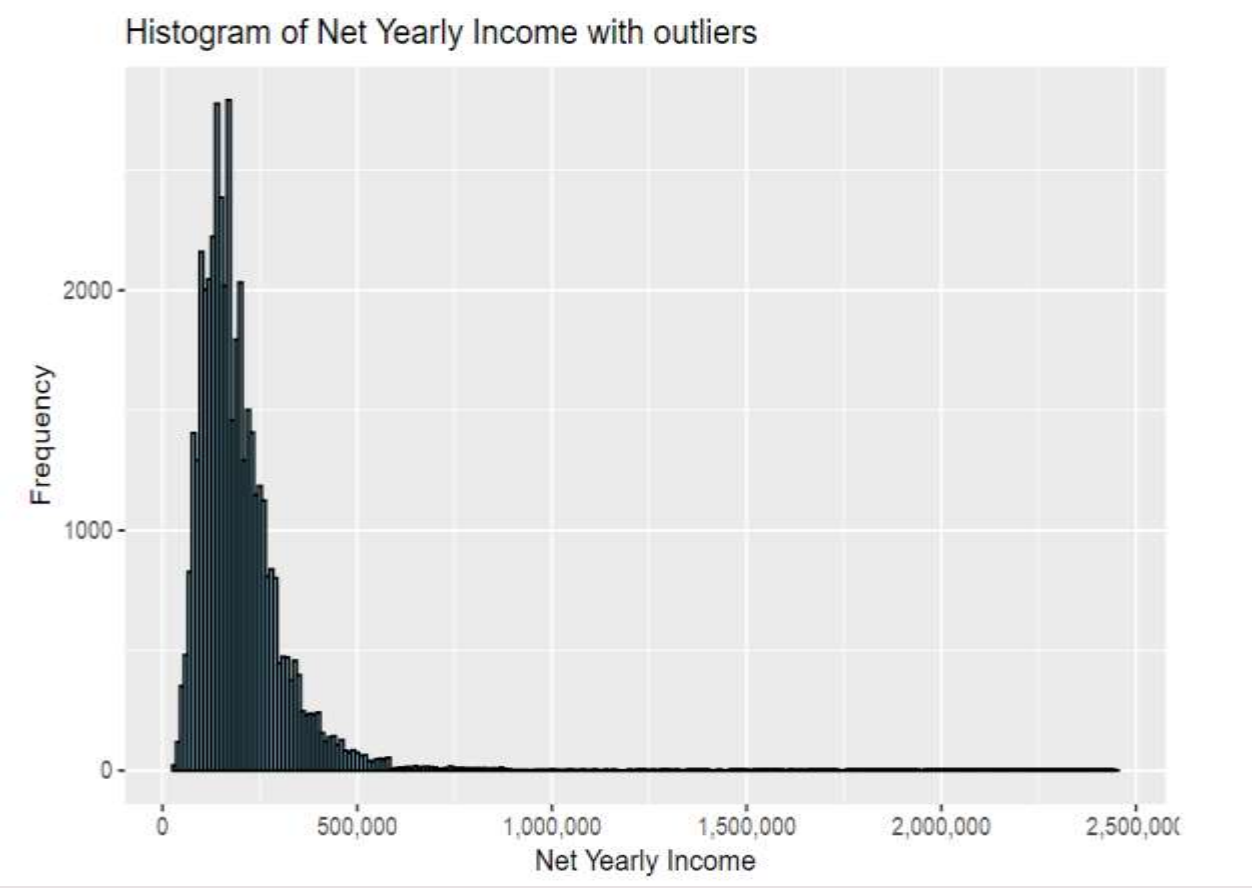
- The dataset for this research proposal is taken from Kaggle of the customers of the bank American Express. The dataset was originally a part of an online hackathon with about 45,528 unique observations. Due to a high number of observations in the training data, we removed any unknown observations!
- The credit limit had potential outliers which were removed after observing the summary statistics. Box plots prior and after removal of outliers are shown below. Note that credit limit potentially has a strong relationship with credit card default.

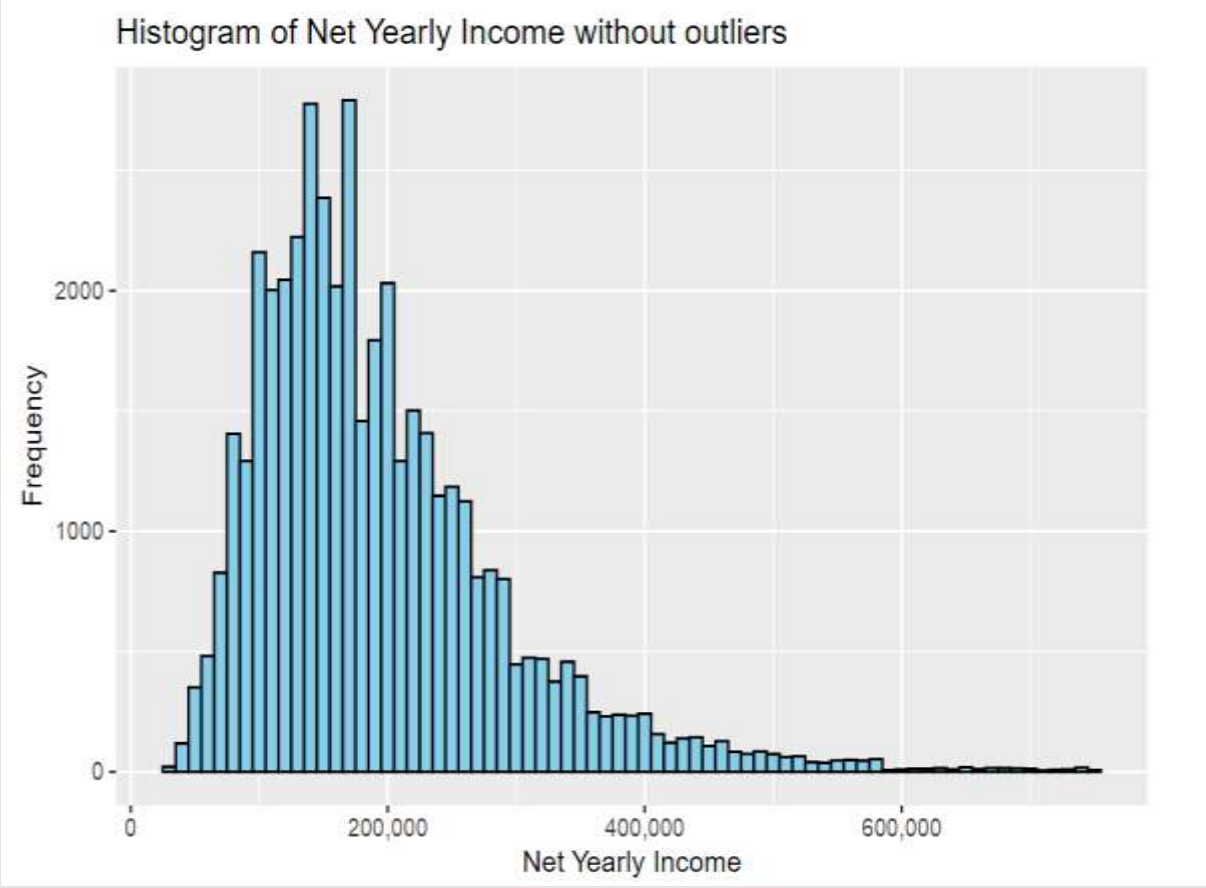




EXPLORATORY DATA ANALYSIS: HIGHLIGHTS

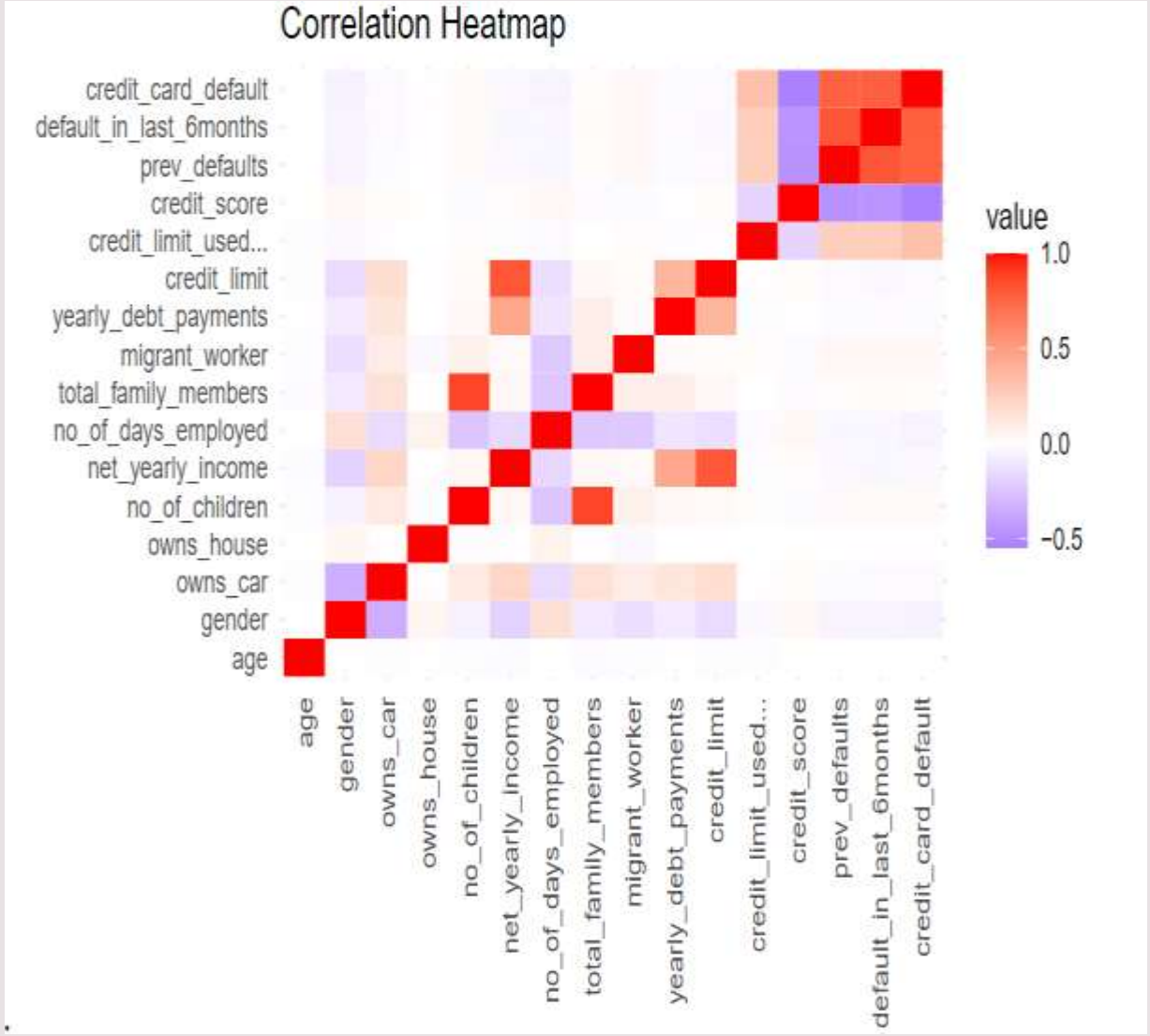
- Another variable that could potentially play a crucial role is net yearly income. Potential outliers were identified and were removed by a cutoff of $750,000 per annum of income. Histograms prior and after removal of outliers is shown below.





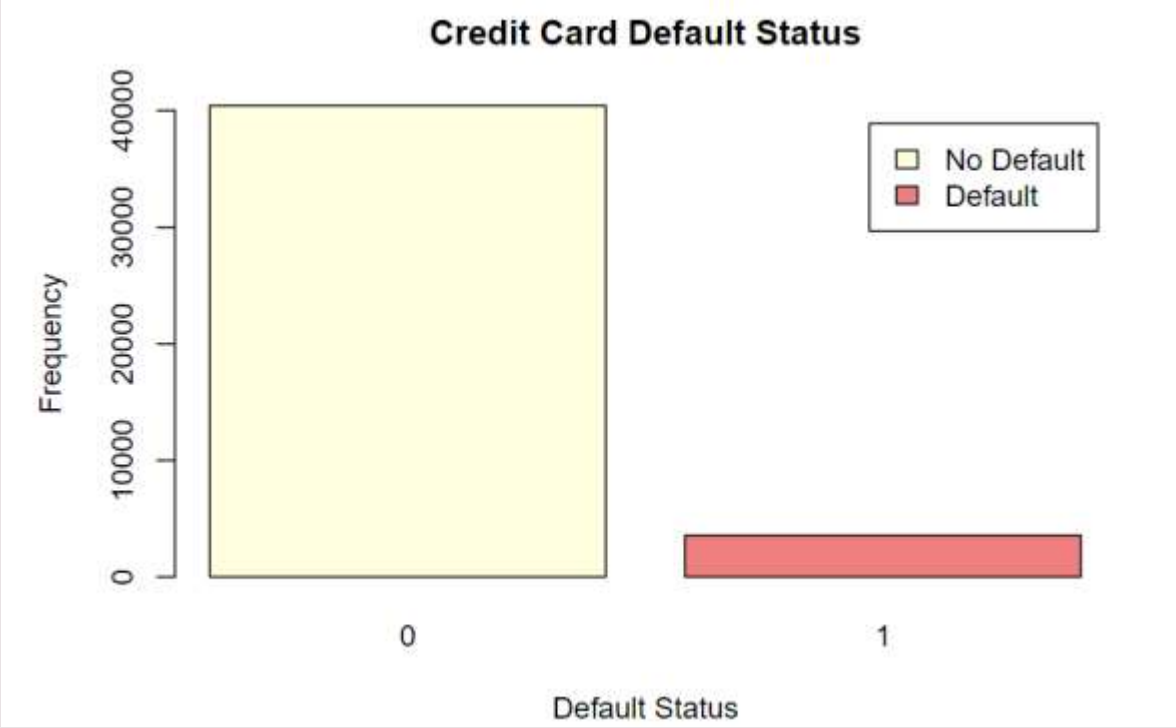
EXPLORATORY DATA ANALYSIS: HIGHLIGHTS

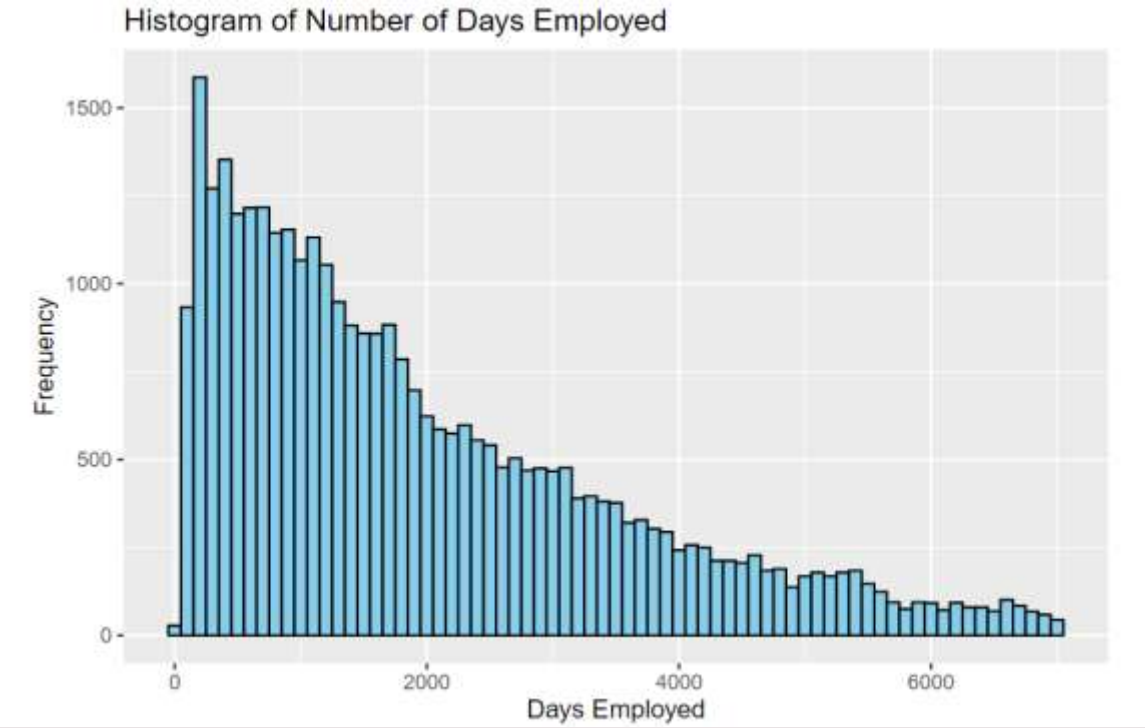
- Correlation between number of children and total family members and between credit limit and net yearly income is observed.
- This may cause issues with multicollinearity.
- We also found potential confounders by observing the heatmap where previous defaults and defaults within 6 months are closely related so we will use only previous defaults.
- Similarly, number of children is a redundant variable since it had a high correlation with total family members.
- **Key Insight:** Negative correlation between credit score and credit default!



GENERALIZED LINEAR MODEL SUITABILITY

- In the context of credit card default prediction, the outcome of interest (whether a customer defaults) is a binary variable (as shown below), making logistic regression, a type of GLM, an ideal choice for the analysis.
- Moreover, my predictor variables including net yearly income (see previous slide) and number of days employed (shown below) do not particularly come from a normal distribution. In fact, they somewhat show a chi-squared distribution which needs to be further studied!





REFERENCES

- Dataset Source: https://www.kaggle.com/datasets/pradip11/amexpert-codelab-2021/data?select=train.csv
- Zhou, X., Zhang, W., & Jiang, Y. (2020). Personal credit default prediction model based on convolution neural network. Mathematical Problems in Engineering, 2020, 1–10. https://doi.org/10.1155/2020/5608392
- Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: a machine learning methodology. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-98361-6
- Sayjadah, Y., Hashem, M., Alotaibi, F., & Kasmiran, K. A. (2018). Credit Card Default Prediction using Machine Learning Techniques. IEEE Xplore. https://doi.org/10.1109/icaccaf.2018.8776802
- Calabrese, R., & Osmetti, S. A. (2011). Generalized Extreme Value Regression for Binary Rare Events Data: an Application to Credit Defaults. RePEc: Research Papers in Economics. https://econpapers.repec.org/RePEc:ucd:wpaper:201120
- Kealhofer, S. (2003). Quantifying Credit Risk I: Default prediction. Financial Analysts Journal, 59(1), 30–44. https://doi.org/10.2469/faj.v59.n1.2501
- Credit Default Prediction Predicting Credit Defaults: A guide for entrepreneurs - FasterCapital. (n.d.). FasterCapital. https://fastercapital.com/content/Credit-Default-Prediction-Predicting-Credit-Defaults--A-Guide-for-Entrepreneurs.html
- Chaudhary, P. (2023, February 21). Credit Default Prediction — Practical tips for successful execution. Medium. https://medium.com/cuenex/credit-default-prediction-practical-tips-for-successful-execution-f62a92ab5df8