Week 7 Report

Tyler Chinn

Colorado State University Global

Dr. Goodwin

January 07, 2022

**Abstract**

The American Express credit default dataset published to Kaggle is investigated to see its potential for predictive power with relationship to available parameters. Additionally, a literature review on credit risk analysis and an overview of financial institution interest in default prediction with machine learning is explored. Exploratory data analysis was conducted in python, using plotly express and the python statistical library. Results of logistic regression, random forest, and light gradient boosted machine models show an accuracy of around 85% in predicting customer default. Data drift was investigated and the holdout test set provided did have parameters that were statistically significant than those in the training and validation sets.

**Section 1. Background and Literature Review**

**Background**

Banks and other institutions are always looking for ways to ensure that the loans they give out are profitable. Additionally, accurately updating information to determine what debt assets still have value is important to stay soluble. For instance, if a bank determines that 10% of its loans are unlikely to be paid back on time or at all it must begin taking actions to protect itself (Adebiyi et al., 2022).

While there are many factors that determine loan risk there are the traditional 5 Cs for credit scoring that include collateral, credit history, capital, condition and capacity. Collateral refers to assets that are sold that could be used to cover the amount of money being used. While the dataset in this study looks at credit card loans, collateral would still refer to any concrete assets that could be liquidated to cover the debt such as property and stocks. Credit history and ratings are the most well known by consumers and refer to the length, amount and repayment of loans of an individual. Capital refers to the loan amount, and in this case could also refer to the amount of credit given to each consumer. Condition refers to the type of usage that the money is for. Loans that are given for investments or business expenses are generally more likely to be paid back than similar amounts for emergency medical expenses. This is because business expenses usually have some rate of return if successful, while other expenses may simply be lost for sure after use. Lastly is capacity, the general ability of the loaner to earn money (Adebiyi et al., 2022).

**Literature Review**

In 2014 the size of outstanding revolving credit was $861 billion. As the financial crisis showed in 2008, banks need to be able to manage both their internal and external risk factors. Credit card accounts are considered revolving lines of credit. Revolving lines of credit are continuously run up and down (as purchases and payments are made) giving lenders a more active ability to manage these portfolios. Financial institutions can save money by freezing credit lines that are predicted to default, with potential savings in the hundreds of millions. There is however risk associated with this method, false positives and subsequent improper freezing of accounts can alienate customers and reduce brand reputation (Butaru et al., 2016).

In order to accomplish this objective, companies must build robust predictive models that accurately predict if an account is at risk of defaulting, and additionally, how much credit should be allocated to each customer. Machine learning models based on a number of frameworks are being continuously investigated and updated. These include linear regression models, decision trees, neural networks, naive bayes and others (Butaru et al., 2016) (Mahbobi et al., 2022). Models built on cutting lines of credit such as those generated by Butaru et al., 2016, are limited to savings on revolving credit lines, and those which can have actively managed amounts of credit. These models cannot predict savings on the portfolio as a whole, nor do they consider additional impacts to customer base, overhead expenses or other expenses.

**Changes in Lending Practices and Effect on Models**

There have been large changes in the past with regards to how credit risk management was handled. In the oldest of systems, credit risk remained on a lender’s balance sheet until repayment or default. In the 1990s loans and risk became routinely reconfigured and sold between financial institutions, and a bank no longer necessarily held the credit risk until maturity (Caouette et al., 1998). Naturally these changes among others have altered the way risk management is handled.

The need for predictive power and modeling however has not changed. For banks and lenders to understand their exact exposure to the market requires an accurate assessment of the ratings of loans. For instance, what classifies junk grade bonds versus investment grade bonds? While credit ratings are outside the scope of this work, it is an example of the continued research into better methods.

**Differences Among Financial Institutions**

A 2016 study using information from financial regulators found that there is a substantial difference in risk factors and susceptibility among banking institutions. Certain banks were much more active and better at managing their credit card risk exposure, making them more resilient to delinquency propagation. The authors further found that even certain labeled high risk portfolios could have fewer delinquency spikes due to successful active risk management by the bank (Butaru et al., 2016).

**Section 2. Project Proposal**

**Problem**

To stay competitive banks must compete with the best models to maximize their profit. Additionally, bank/loan members that deal with customers must increasingly make data driven decisions in order to stay competitive in the marketplace. Therefore, both democratizing data across an organization (dashboards to show loan performance) and better predictive power models are needed.

**Purpose Statement**

The purpose of this research project is to demonstrate several technologies that American Express can leverage to both examine their large amounts of customer data and help employees across the organization take advantage of the results. This will include visualization and dashboarding analytic software such as Tableau, machine learning and predictive analytics with python (and packages).

**Deliverables**

The goal is to generate the strongest predictive model possible with the use of several machine learning methods and data preparation techniques. This will likely involve popular tools such as python, pandas, matplotlib and sklearn.. While American Express did not provide company information such as its previous and current predictive power with regards to the data, the baseline of the default class percentage will be used. That is to say the model would have to beat a model that simply guesses the majority class. Any model that does worse than this would have little value, so this serves as an important baseline.

**Dataset**

This dataset was chosen because it was a readily available credit dataset that was for public use. While the competition to which the dataset was released has since finished, the company American Express was seeking real value in this dataset, and as such provided millions of entries from presumably their own database (or from a database available to financial institutions) (American Express, 2022). For this reason it was chosen to examine machine learning models and analytics for the financial industry.

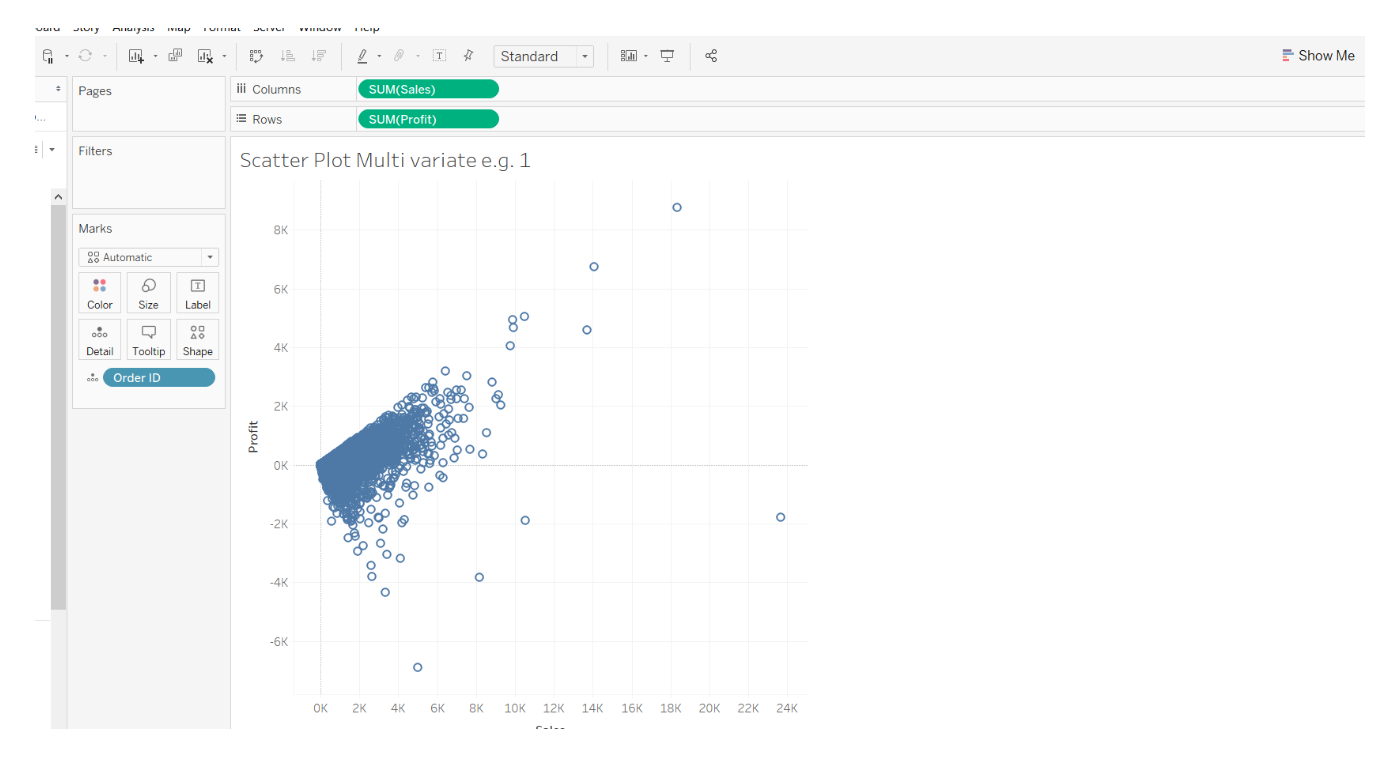
**Limitations**

The data is fully anonymized so variables will have to be referred to by their ambiguous column names (although there is some information on data types related to financial instances). In many cases this is not unusual to prevent privacy leaks by employees.

**Techniques for Analysis**

There are several tools that are used to analyze the dataset. Tableau is used to generate dashboards to show the number of default targets and how it correlates to the dataset variables.

Figure 1. Tableau EDA Example



Note: Adapted from PluralSight https://www.pluralsight.com/guides/exploratory-data-analysis-with-tableau

Python is also used in order to analyze the data. The following figures were generated from the dataset.

Figure 2. First Rows of Dataset

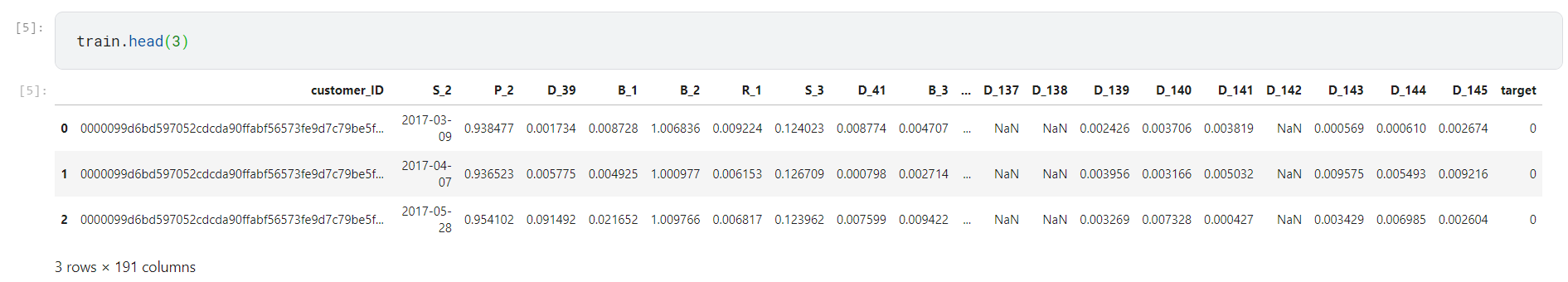
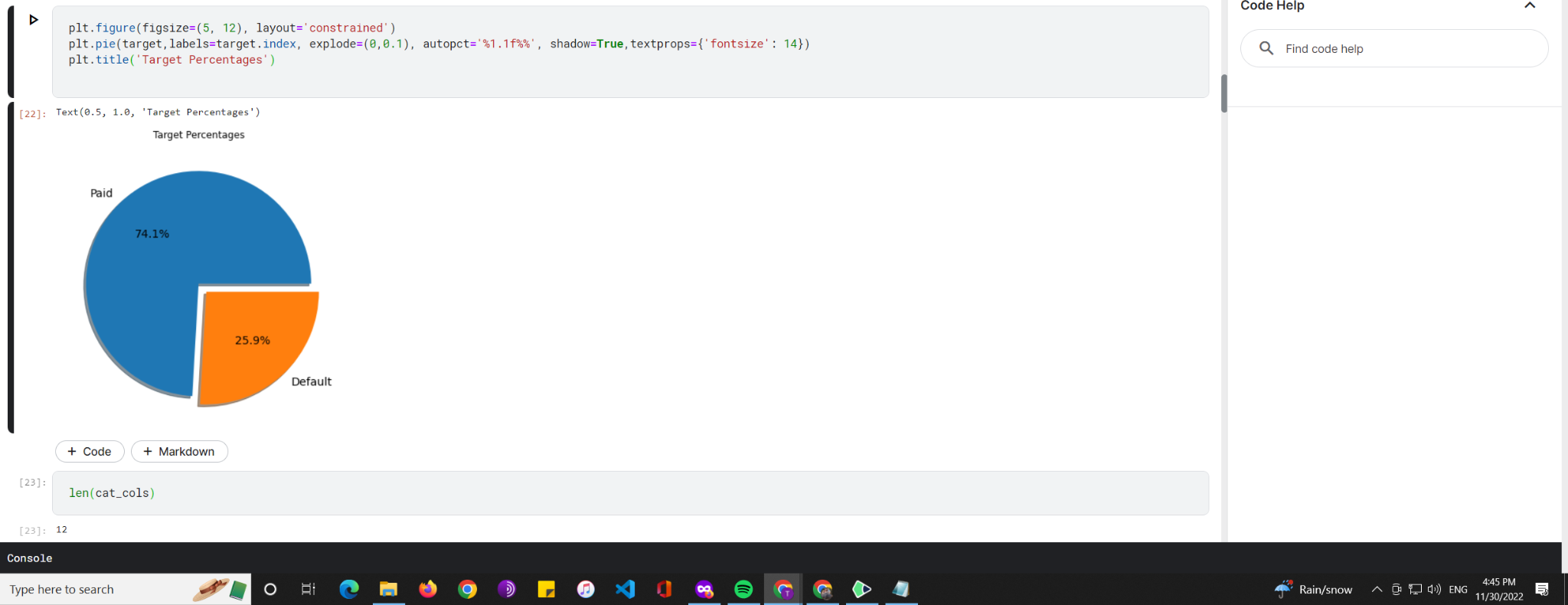
Using packages such as matplotlib and plotly express, interactive figures can be generated on all variables in an efficient way.

Figure 3. Pie Chart of Target Variable (Default vs Paid)



Python is also used in order to create a machine learning model. There are many different models, parameter requirements and the time to generate each is costly based on the dataset size.

Tree models have interpretability preferences compared to neural networks. There is sometimes an unfortunate tradeoff between explainability of a machine learning model and absolute predictive power. This is not necessarily the case however, and multiple models can be used to try and gain the most of both (Goyal & Kaur, 2016). Additionally, neural networks do not handle missing variables so as the data

**Section 3. Methodology**

**Description of Variables**

As mentioned in the project proposal there are several different categories of variables in the dataset. These have been anonymized both to protect American Expresses proprietary data collection and advantages, as well as protect consumer privacy. The categories of data are spend variables, payment variables, delinquency variables, balance variables, and risk variables. More is discussed such as the Dtype of each variable, and the count of non null variables in the visual model of the data section.   
 While there is no additional information there are several general concepts to credit loans in the industry. Credit history and ratings are the most well known by consumers and refer to the length, amount and repayment of loans of an individual. Capital refers to the loan amount, and in this case could also refer to the amount of credit given to each consumer. Condition refers to the type of usage that the money is for. Loans that are given for investments or business expenses are generally more likely to be paid back than similar amounts for emergency medical expenses. This is because business expenses usually have some rate of return if successful, while other expenses may simply be lost for sure after use. Lastly is capacity, the general ability of the loaner to earn money (Adebiyi et al., 2022).

The goal of the bank is to determine the percentage chance that a customer does not pay back their credit card balance based on their customer profile from the past 18 months. If the customer does not pay the due amount in 120 days after their last statement date it is considered a default event.

The official metrics that will be used in scoring the machine learning model portion of this project is provided by American Express as follows: “The evaluation metric, M, for this competition is the mean of two measures of rank ordering: Normalized Gini Coefficient, G, and default rate captured at 4%, D



The default rate captured at 4% is the percentage of the positive labels (defaults) captured within the highest-ranked 4% of the predictions, and represents a Sensitivity/Recall statistic.

For both of the sub-metrics G and D, the negative labels are given a weight of 20 to adjust for downsampling. This metric has a maximum value of 1.0.” (American Express - default prediction).

American Express is the largest card issuer in the world, and has stated that this dataset it has provided is of industry scale. Additionally, the information has been presplit into a reserved testing dataset that the company uses to judge predictive performance (to prevent intentional data leakage). The company also is interested in any tools that create the most powerful model (American Express - default prediction).

**Research Questions**

Is there a relationship between the variables and the target outcome? This is ultimately the goal of the project to determine how the attributes can be used to determine loan default rates. Additionally, a follow up to this question could be to see what variables are not relevant to building the model, and if these even need to be continuously collected. Improving predictive models and assessing their performance is an ongoing issue, and updates are continuously necessary due to shifts in the market (Butaru et al., 2016).

Do the train and test datasets come from the same distribution? This is more a test to determine if the groups are different or not from the test and the training dataset. If the populations are different this could be used to examine if there is missing data, biased data due to sampling techniques etc. that are being used for the test set.

Additionally, it is important to check if there was any artificial bias that might have been introduced in the competition. As can be seen from the last statement distribution for every customer, the most recent statements for the test data are different than from the training data. This is natural from a setup point of view as a model that cannot infer properly from this new data is likely overfitting. Despite this it is useful to know what type and extent of data drift can be expected in the test set. If there is significant data drift over this period, confidence intervals may have to be adjusted to be more representative of expected results. Likely when this model would actually get deployed, it would be retrained on the most recent data, as is common with time series analysis.

**List of Hypothesis for Capstone Project**

**Q1 H0**

There is no relationship between the predictor variables and the target variable.

**Q1 Ha**

There is a relationship between the variable and the target variable

**Q2 Test of Relationship for Categorical Variables H0**

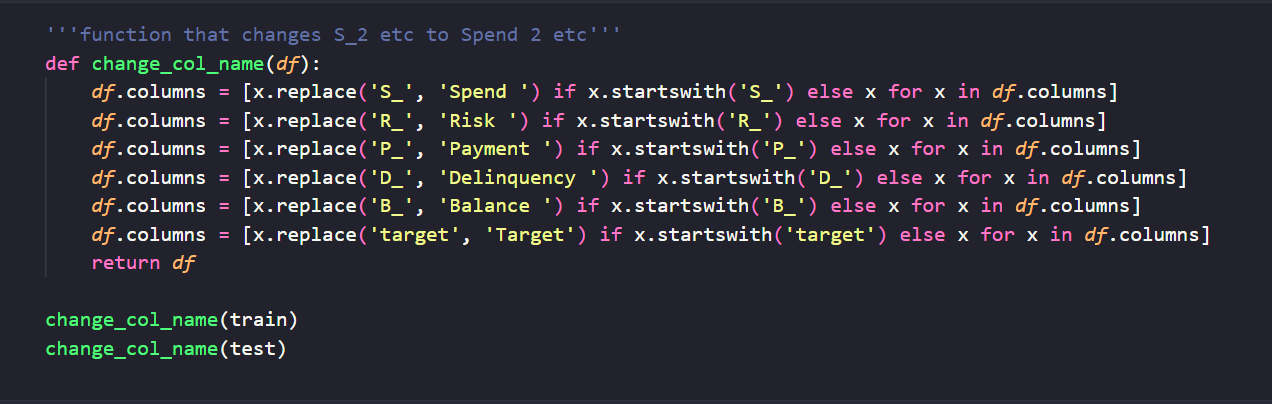
The train and test datasets have the same distribution of categorical variables. That is to say there is no statistically significant difference in the distribution of the categorical variable between the train and test set.

**Q2 Ha**

The train and test datasets have different distributions of categorical variables.

**Section 3. Methods**

Figure 4. Sample python code for data wrangling.



Python was used for the majority of the investigation as it has the best performance when working with large datasets. That being said Tableau offers an easy to use interface when trying to create certain graphics.

**Statistical Methods**

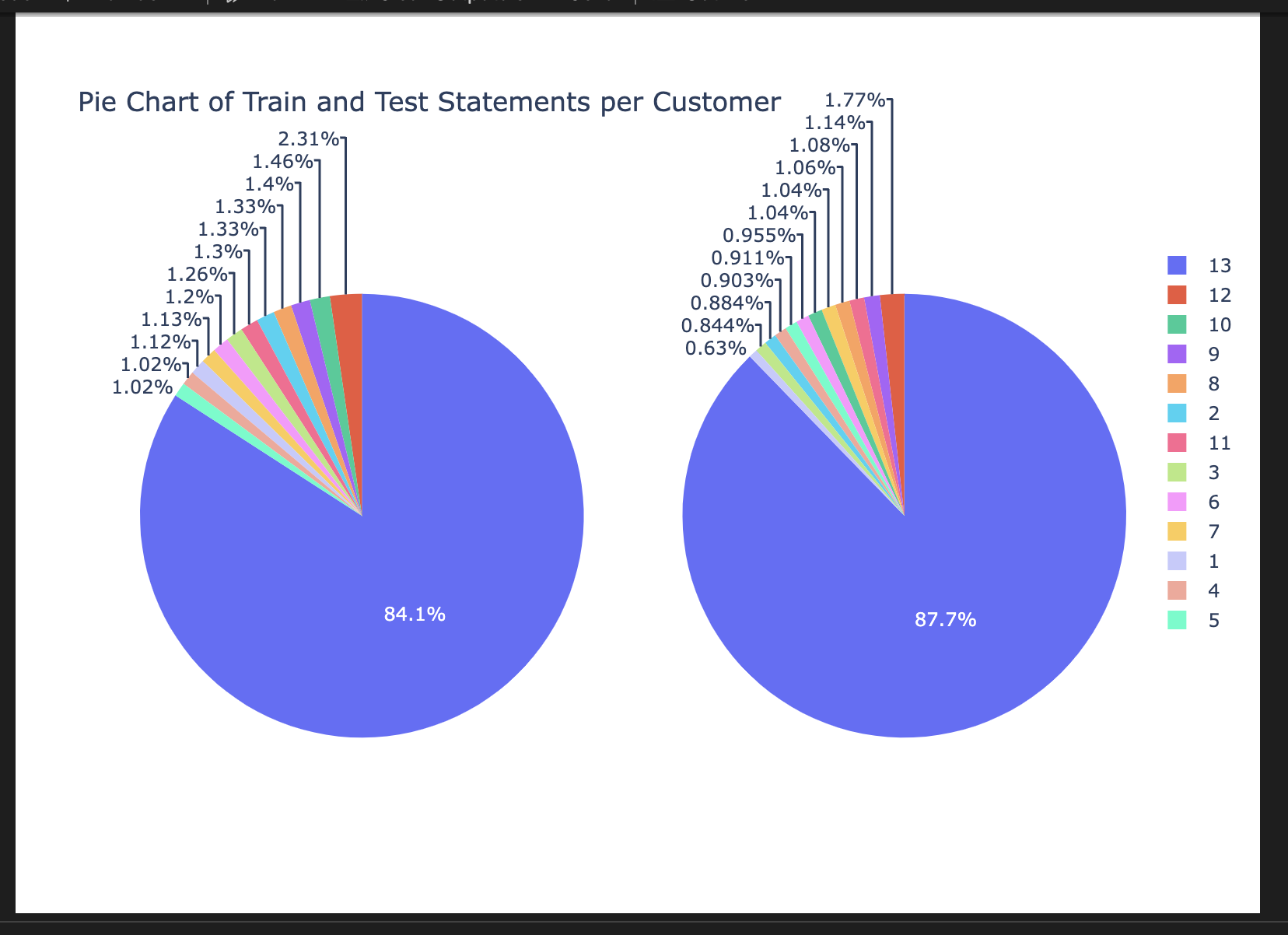
For testing the first hypothesis if there is a relationship between the target and outcome there are several tests that can be used. To start a Pearson or Spearman correlation matrix can be created. It will display how strongly correlated variables are with the target variable. Additionally, this may reveal any strong correlation between other variables. Spearman rank correlation coefficient is also a non parametric option. After this test is conducted the end goal of the project is to build a predictive model, and this model will hopefully show statistically significant improvements over random guessing. The probability of this being true will also be tested (Goyal & Kaur, 2016). Lastly a Chi-Squared test will show how the model is performing in terms of misclassification of default. Other metrics such as accuracy and f score will also be used. It should be noted that the dataset has been upsampled already for the target classification to make the training data more balanced. The weighting used in the competition takes this factor into consideration.

For testing the hypothesis if the populations come from the same distribution a t-test or a Welch’s t-test can be used. A Welch’s test is often used both when the standard deviations are different and when sample sizes are uneven. This will show if the samples are coming from the same population within a certain degree of certainty. In this case, it is important to consider if the populations given by American Express are different from the test set, or if the populations have deviated due to some external factor.

**Results of Exploratory Data Analysis**

The results of the statistical analysis show that there are some differences with the test and train data distribution. To start there is a clear difference on the distribution of the last statement that each customer id has in the dataset. A graph of this solution is provided.

Figure 5. Examination of Number of Statements per Customer



Looking more specifically at the results of distribution for delinquency variables 63 and 64 it is also clear that the samples do not come from the same population.

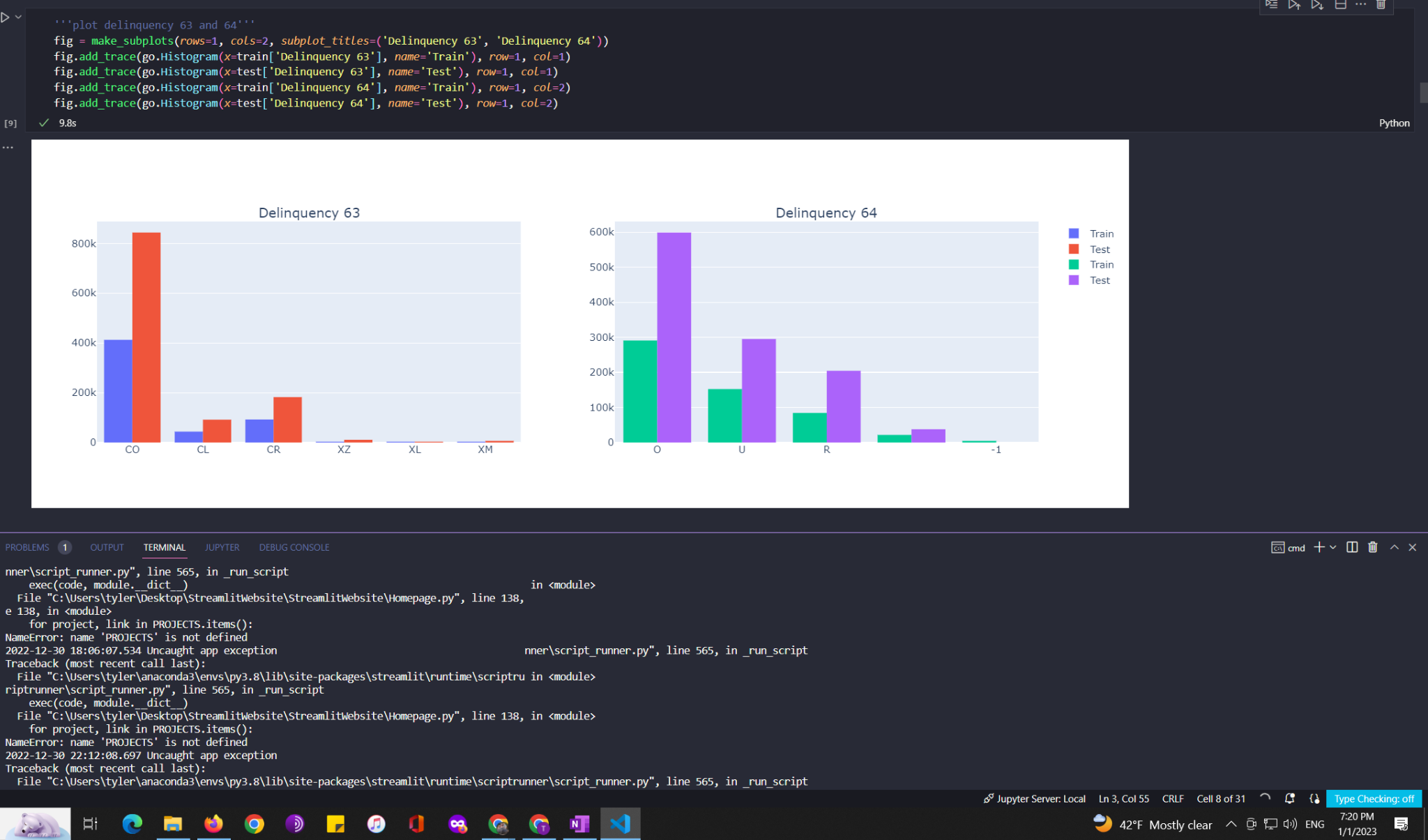
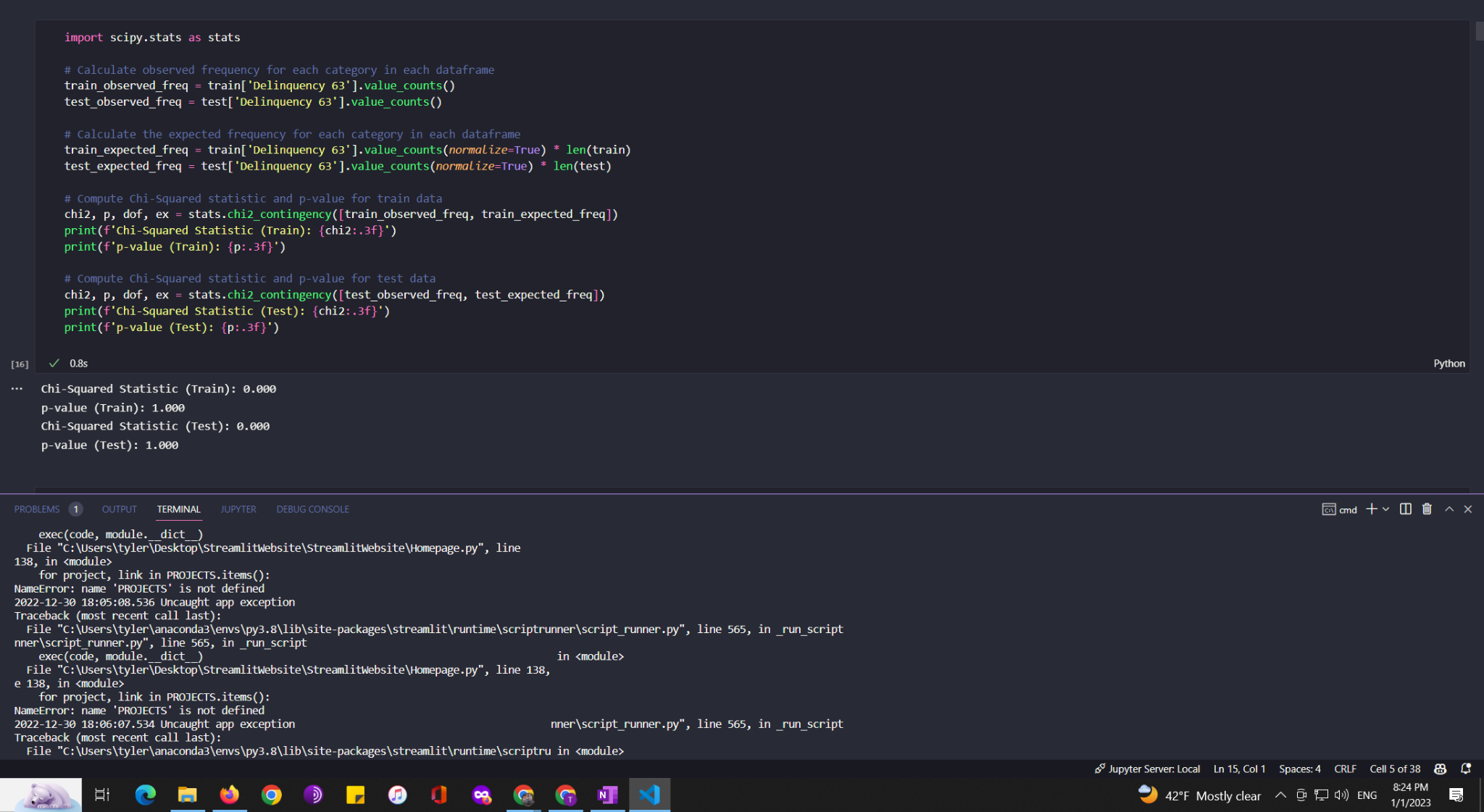
Figure 6.

Figure 7. Results of Chi Square Test



The results of the chi square test fails to reject the null hypothesis that the groups are independent. Looking at the histogram for variables 63, and 64 there does appear to be a significant difference in the distribution of variables. Due to this fact, it can be concluded that there are certainly statistically significant differences between the train and test groups (Géron, 2022). Combined with the difference in statement time spread, it can be concluded that there is a reasonable difference between the two groups. In this instance where data drift can be expected, it is important for any machine learning model to not overfit the training dataset. Additionally, based on this knowledge it would be recommended for the company to update its credit default model frequently.

This data drift is certainly not uncommon in the financial industry or other consumer industries as well. There can be not only economic changes, but changes in consumer behavior that could alter the risk of default from a perceived couple of variables.

**Results of Machine Learning Models**

Figure 8. Logistic Regression Model

The results of a Logistic Regression Model are shown above. Using only numeric columns with non missing values the model achieves an accuracy of 0.85.

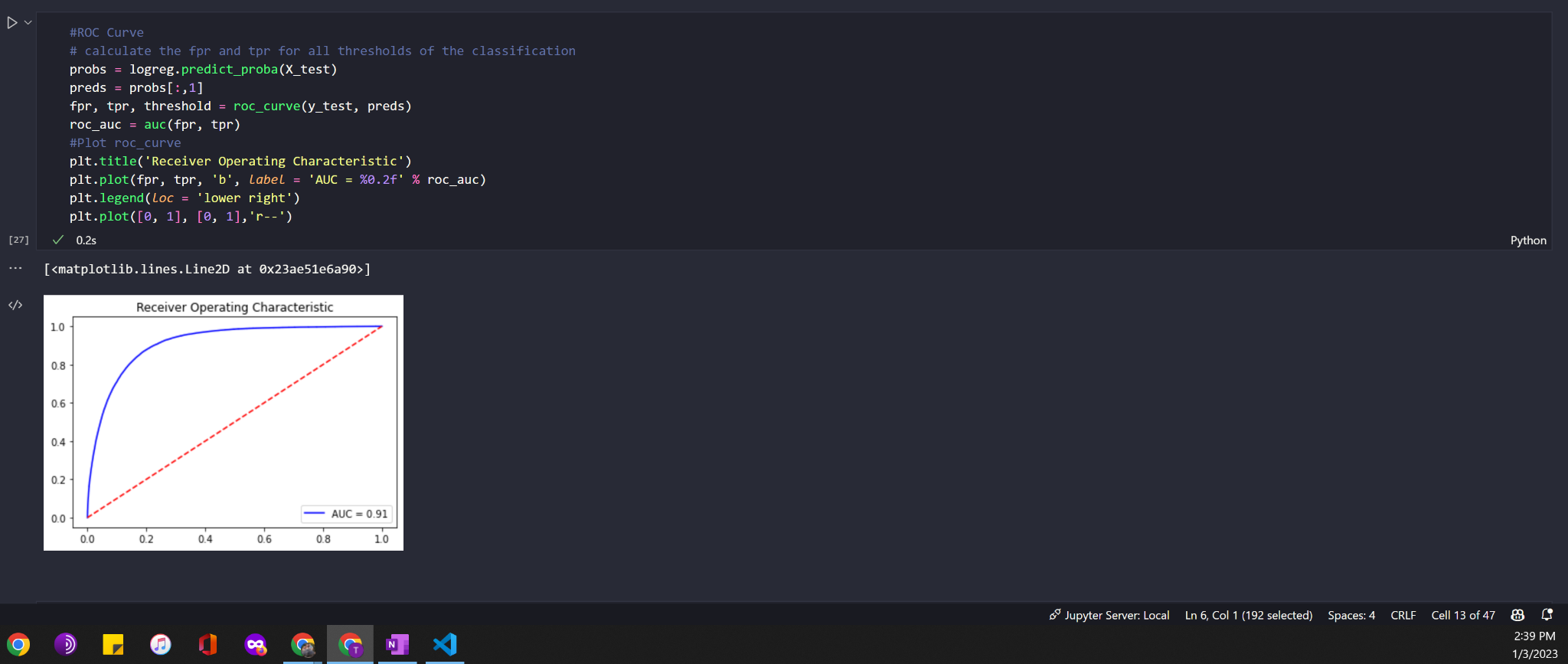
Figure 9. Results of Logistic Regression Model

Figure 10. Confusion Matrix of Logistic Regression

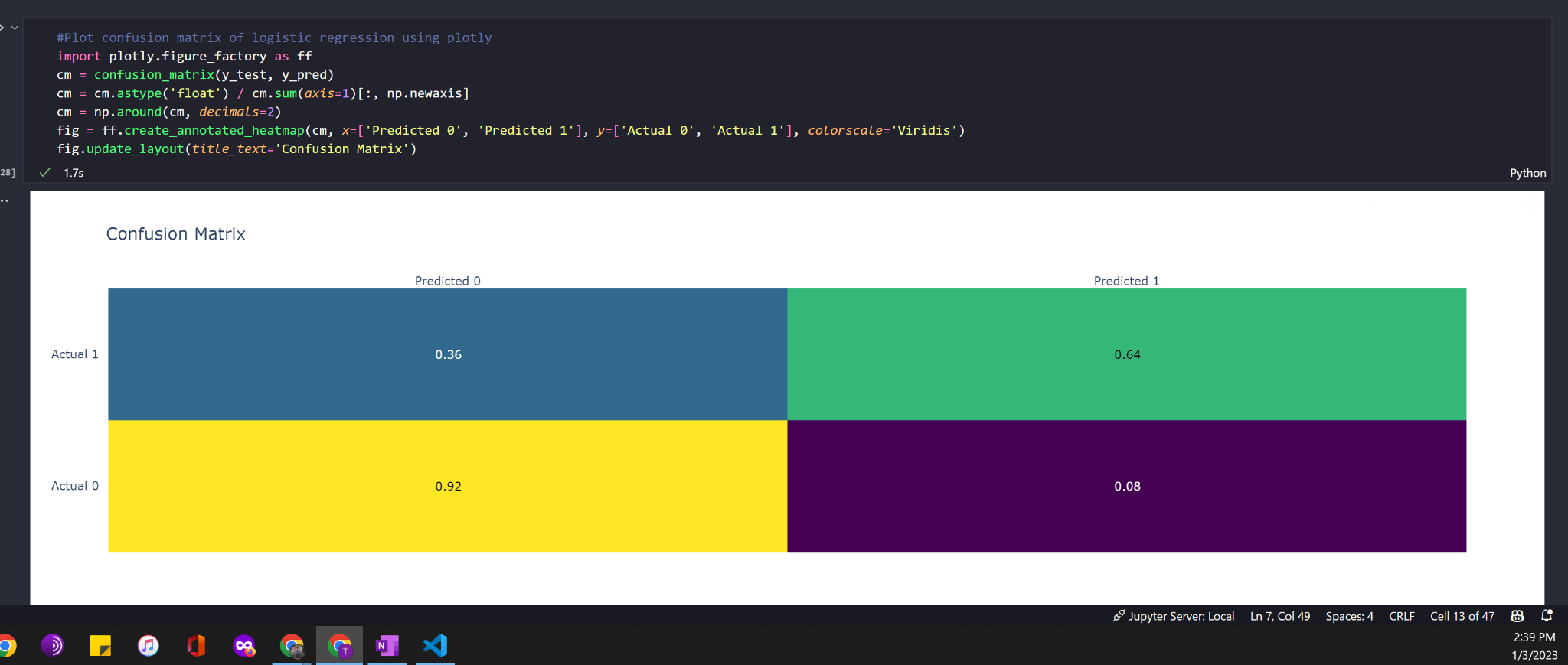


Figure 11. Effects of Standard Scaling on Model Performance

 The above models outperformed a baseline performance of around 75% accuracy with an accuracy of 85% (baseline performance being predicting the default class). While sklearn models do not natively support pvalue statistics for predictive power the python stats version of logistic regression does. This model confirms that there is indeed a statistically significant p<0.05 relationship between the variables chosen with the target default variable. Chi-Square results are provided as well.

**Section 4. Limitations**

There are several limitations with this project and scope of work. Due to anonymization and added noise to the data it is not possible to determine if there are any specific ethical considerations from the data. An additional limitation from a business applications side is that there is no information on how this data was collected. It could be a non representative sample, or it may not properly represent the actual organizational conditions for the credit loans. One example would be if the delinquency or risk variables are a subjective rating from individual bank employees. If so, assuming that there is no potential geographical or other biases in the data from this process would be incorrect. While some additional investigation could also be done on separation of customers into groups through unsupervised techniques, there is no way to determine what these groups would mean without additional variable information. For instance if the goal was to separate based on customer type (working professionals, manufacturing, unemployed etc.) this would be vague at best due to the anonymization.

Another limitation due to cost constraints there were severe limitations on the amount of working ram available for training and exploratory data analysis. As mentioned in the methods section, the data was cast to lower forms (in certain columns there was no information loss as there were integers being used as floats) which reduces information in the training set. While some machine learning models allow batch processing, LGBM only allows for distributed processing across multiple machines for increasing beyond one machine's ram constraints. It is therefore required to work with this loss of information.

There is still a need to use business rules to implement any predictions or knowledge gained from this project. For instance what percentage of credit should be limited on what risk percentage? Does certain customer retention outweigh much higher risk of default? These will need to be discussed, defined and implemented.

Lastly, like any model there is a risk of data drift, and finance is likely a field that experiences a good amount of this due to changing macroeconomic conditions and government regulation changes. Careful monitoring for this will be necessary.

**Section 5. Ethical Considerations**

**Privacy and Security Concerns**

As previously mentioned in the project proposal there are a number of ethical considerations that need to be accounted for. A quick literature review reveals some of the common problems associated with credit card loan analysis and ethics. While the data has been anonymized there is the possibility that a leak of information from American Express could also potentially compromise this data in the future, or that it could be linked to individuals. However, due to the nature of credit history, loans etc. most of the information on defaults is already available to financial institutions American Express - default prediction (Kaggle, 2022.). On the level specifically related to this body of work, while there is potential for ethical abuse in the field, there was no specific security measures that were needed for protecting the dataset.

**Ethical Considerations**

A journal article related to disparity in results for healthcare discusses some of the major risks with automation. Some risks of developing these models include an overreliance on automation with a lack of interdisciplinary approach to decision making. Additionally, algorithms can be based on biased data, good datasets require testing and developing models in socioeconomic diverse systems. Lastly, feedback loops need to be monitored. In the case of finance this could be that low income households are denied credit, and therefore these households' incomes fall in comparison to peers. One way to address this is to input ethical values into models, even at the cost of some predictive power (Gianfrancesco et al., 2018). Machine learning algorithms are being rapidly adopted in the finance industry as well. Like the study related to healthcare in the credit market (including housing loans, credit cards and rates) the models have shown a tendency to increase disparity among ethnic groups (Fuster et al., 2022).

Another ethical consideration that must be considered when making machine learning models and predictive analytics is if there is bias in the models that could discriminate against certain groups of people. In fact the bias may not even need the actual feature in the data model if there are strongly correlated features. For example if there is information within the features that could determine someone's health or if they were disabled to a high degree of certainty, a model could actually be using this to make decisions. This is also why certain black box techniques are unpopular in areas that are required to show no bias against protected groups of people.

Ultimately as this is an anonymous data set with the variable names taken out to protect privacy and company information this is not a specific concern for the analysis. If the information becomes available at some point in the future there are certain techniques and packages that can be used to test models for unfair bias.

Another consideration that should be taken into consideration is how this information is going to be used. Studies show that credit card debt and overhead can be a great stressor on wellbeing. While loan repayment and customer values hopefully align, it should be noted that in certain situations maximum profit comes at a potential cost to the customers (Norvilitis et al., 2006). Simply having a system that tries to predict loan repayment is unlikely to take this into consideration, and amore holistic approach may still be advisable.

**Conclusions**

The above work has investigated the connection between anonymized credit data from millions of customer data points.

Exploratory data analysis on the American Express dataset showed that there are some time series component differences between the train and test set. Additionally, the results of exploratory data analysis suggest data drift over the time series.

**Future Work**

There are several areas that would benefit from future modeling. First there was minimal hyperparameter tuning. Hyperparameter tuning can generally result in better predictive power of models. With a system with more RAM and GPU memory, more of the training data could be utilized. This would be needed for running the full set on a neural network, as it stands only a small percentage was utilized. Generally more training data can improve model performance, although there is a risk of overfitting. If this is not an option there are models that handle distributed processing well (Géron, 2022). Lastly with this information a public Tableau dashboard could be generated on top of the data analysis conducted within Tableau. Tableau was used for quick data investigation, but it could now be used to show the feature importance of the predictor variables. Additionally, it could also be used to visualize any data drifts that could require monitoring.

**References**

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in

machine learning algorithms using electronic health record data. JAMA internal medicine, 178(11), 1544-1547.

Fuster, A., Goldsmith‐Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal?

The effects of machine learning on credit markets. The Journal of Finance, 77(1), 5-47.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in

the credit card industry. Journal of Banking & Finance, 72, 218-239.

Goyal, A., & Kaur, R. (2016). A survey on ensemble model for loan prediction. International Journal of

Engineering Trends and Applications (IJETA), 3(1), 32-37.

American Express - default prediction. Kaggle. (n.d.). Retrieved November 18, 2022, from

https://www.kaggle.com/competitions/amex-default-prediction/overview/evaluation

Machi, L. MacEvoy, B. T. (2016). The literature review – Six steps to success. (3rd ed.). Corwin. ISBN: 9781506336244

ADEBIYI, M. O., ADEOYE, O. O., OGUNDOKUN, R. O., OKESOLA, J. O., & ADEBIYI, A.

A. (2022). SECURED LOAN PREDICTION SYSTEM USING ARTIFICIAL NEURAL NETWORK. *Journal of Engineering Science and Technology*, *17*(2), 0854-0873.

O’Leary, Z. (2017). *The essential guide to doing your research project* (3rd ed.). Sage

Publishing. ISBN: 9781473952089

Norvilitis, J. M., Merwin, M. M., Osberg, T. M., Roehling, P. V., Young, P., & Kamas, M. M.

(2006). Personality factors, money attitudes, financial knowledge, and credit‐card debt in college students 1. *Journal of applied social psychology*, *36*(6), 1395-1413.

Caouette, J. B., Caouette, J. B., Altman, E. I., & Narayanan, P. (1998). Managing credit risk: the next

great financial challenge.

Mahbobi, M., Kimiagari, S., & Vasudevan, M. (2021). Credit risk classification: an integrated predictive

accuracy algorithm using artificial and deep neural networks. *Annals of Operations Research*, 1-29.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments.

Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly

Media,

Inc.".