MSc Digital Innovation 2021-22 (Summer Term)

Module: MIS 41230 Machine Learning for Business

Module Co-ordinator: Dr Linus Wunderlich / Dr Yossi Lichtenstein

To default or not to default: Assessing the likelihood of late repayments in credit card loan applications using binary and multiclass machine learning techniques.

Kate Nolan

Student Number: 17326336

Kate.nolan1@ucdconnect.ie

Nathan Summers

Student Number: 21206391

Nathan.summers@ucdconnect.ie

Conor Fitzpatrick

Student Number: 17395433

Conor.fitzpatrick3@ucdconnect.ie

Declaration of Authorship

We declare that all materials included in this research report are the result of our own work and that due acknowledgement has been given in the bibliography and in references to all sources be they printed, electronic, or personal.

Kate, Nathan, Conor

Contents

1. Executive Summary

Machine learning has firmly cemented its role as a cornerstone in business, its multifarious capabilities continue to be implemented and many more use cases discovered. The future of strategy and decision making relies on the use of machine learning, allowing companies to make informed decisions based on data and models. The dataset selected for this project is "credit card approval prediction" acquired via Kaggle. Kaggle hosts a community of data-science enthusiasts and houses a huge repository of community published data. Using the data, this project aims to train a model which can determine whether applicants are likely to pay their credit card loans back on time, whether late repayments are likely to eventually lead to full repayments. This analysis has been divided into binary and multi-class problems, therefore several models will be necessary. Hence, logistic regressions, random forests, support vector machines and finally neural networks will be used. A detailed description and results of each analysis follows in this report. This extensive study has provided many interesting findings and patterns.

2. Introduction and Literature Review

Today's business environment is complex and banks, face increasing competition and change. With the introduction of many online banks and credit providers, traditional banks must ensure that they avoid losses and defaulting loans. Since the 2008 global financial crisis, risk management in banks has gained increased prominence, with heightened focus around risk detection, measurement and management (Sharma and Maddulety, 2019). Machine learning has been identified as one of the technologies with important implications for risk management in the banking industry (Sharma and Maddulety, 2019). Machine learning enables the building of more accurate risk models by identifying complex, non-linear patterns within large datasets.

This research will focus on credit card approval. Credit card approval is based on credit scores. This uses personal information and data submitted by the applicant to predict the probability of future defaults and credit card borrowings. Credit scores can objectively quantify the magnitude of risk, and therefore determine whether an applicant will receive a credit card or not (Kaggle, 2020).

This analysis aims to reach several conclusions using machine learning models, the first is whether a loan should be given. Secondly, whether the applicant will repay their loan. Finally, how late are they likely to be in repaying their loan. To determine the above, various machine learning models will be used. These have been selected based on their suitability to the dataset, the problems this research aims to solve and based on an extensive model testing process. The research problems have been divided into binary and multi class problems. For the binary problems, the models proposed are logistic regression, random forest, and support vector machines (SVM). For the multi class problems, the models in use are random forest, artificial neural network and finally SVM. These will be explained below.

The binary problem is whether an applicant is likely to ever repay their loans more than two months overdue. A bad applicant has been defined as someone who has been more than two months late on their repayments in their credit history, hence, the model will predict whether someone is a good or bad applicant based on this threshold. This definition was constructed based on the findings from Kim, Cho and Ryu (2018) which defines credit card loan delinquency as anything beyond three months late on repayments. To solve this problem, logistic regression will be used. Regressions are versatile, they can measure associations, predict outcomes and control for confounding variable effects (Stoltzfus, 2011). Logistic regressions are an efficient way to analyse the effect of a group of independent variables on a binary outcome by quantifying each independent variable's unique contribution (Stoltzfus, 2011). Additionally, random forests will be used. Random forests are much like decision trees, they are an ensemble of trees, making them more accurate (Breiman, 2001). A key advantage of random forests is their ability to capture non-linear association of patterns, this will be essential to identify applicants who have paid late on several, non-consecutive occasions (Kühnlein et al., 2014). Finally, SVM will be used. SVM can be used for forecasting of non-linear components (Zhang et al., 2016). SVM is a powerful method for building a classifier, it aims to create a boundary between two classes that enables the prediction of labels from one or more feature vectors (Huang et al., 2018). The data requires the construction of a label, hence SVM will be useful in the creation of the label.

The multiclass problem is how late people are likely to repay their loans. To access this problem random forests and SVM will be used. Additionally, deep neural networks will be used. Similar to SVM and random forests, neural networks can be used in the forecasting of non-linear components (Zhang et al., 2016).

The above machine learning models were chosen based on the specific criteria and limitations presented by the dataset and the business case. Having introduced and identified relevant literature and models proposed by machine learning scholars, this report will now introduce the specific dataset in use and the business case that prompted the research problems.

3. Dataset

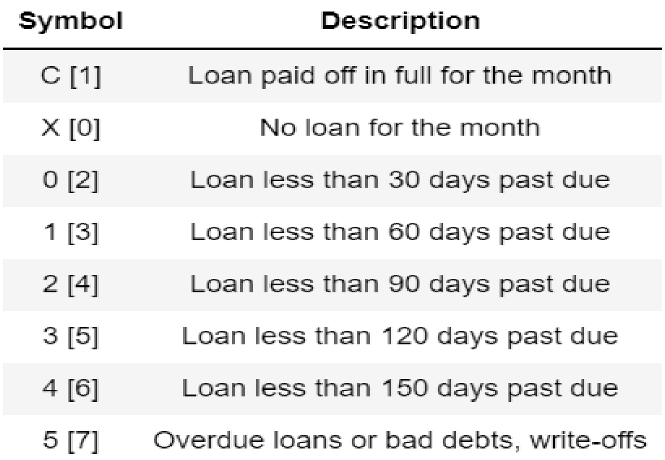
The dataset that will be utilised in this project is entitled "Credit Card Approval Prediction". The dataset can be found on Kaggle, the open-source data platform, and was uploaded in 2020. The dataset is split into two tables, both of which detail information about credit card applications from the perspective of a bank. Combined, this dataset contains more than 1,500,000 data points, however, there are 36,457 unique entries shared between both tables which will be used for the purposes of this analysis.

The first table, entitled application_record.csv, contains information about customers' socio-economic background. This is shown through 17 potential features in the application records table such as income, occupation, number of children and education. A full list of the features included in this table can be found in Table 1.

Feature name	Explanation
ID	Client number
CODE_GENDER	Gender
FLAG_OWN_CAR	Is there a car
FLAG_OWN_REALTY	Is there a property
CNT_CHILDREN	Number of children
AMT_INCOME_TOTAL	Annual income
NAME_INCOME_TYPE	Income category
NAME_EDUCATION_TYPE	Education level
NAME_FAMILY_STATUS	Marital status
NAME_HOUSING_TYPE	Way of living
DAYS_BIRTH	Birthday
DAYS_EMPLOYED	Start date of employment
FLAG_MOBIL	Is there a mobile phone
FLAG_WORK_PHONE	Is there a work phone
FLAG_PHONE	Is there a phone
FLAG_EMAIL	Is there an email
OCCUPATION_TYPE	Occupation
CNT_FAM_MEMBERS	Family size

Table 1: Dataset feature names and explanations

The second table, entitled credit_record.csv, shows the loan repayment status of the candidates and the months balance, which shows the recorded month of the repayment status. The status of the candidate's repayment ranges from numerical values of 0-5 and alphabetical values of C and X. C indicates that an individual has paid off their loan in full for the month, while X indicates that no loan has been taken out for the month. A detailed explanation of the numerical values can be seen in Figure 1. Status symbols are encoded with the value in square brackets when appropriate.



Symbol	Description
C [1]	Loan paid off in full for the month
X [0]	No loan for the month
0 [2]	Loan less than 30 days past due
1 [3]	Loan less than 60 days past due
2 [4]	Loan less than 90 days past due
3 [5]	Loan less than 120 days past due
4 [6]	Loan less than 150 days past due
5 [7]	Overdue loans or bad debts, write-offs

Figure 1: Description of Numerical Values

Both the application-records table and the credit record table contain the "ID" feature which assigns a numerical value to each candidate to hide their identity. This ID feature will operate as a link between the two tables. However, since credit_record.csv contains data for several months for each individual client, there are several copies of each client ID number within the table. Furthermore, there are several clients on one CSV file that do not appear in the other. Considering this, a list of unique ID numbers shared across both CSV files was created and stored in shared_id_list.

4. Business Case

The business case at hand takes the perspective of a bank who are dealing with credit card holding customers. The bank aims to assess whether applicants are likely to pay their credit card loans back on time, whether late repayments are likely to eventually lead to full repayments.

The credit card business model emerged in California in the 1950's by Bank of America. The aim of this venture was to facilitate their customers in the purchasing of goods from a variety of local merchants on credit, allowing the banks to profit both off the credit card holders and the merchants that use their service (Gendal, 2014). As the credit card industry expanded, this business model stayed the same with banks earning revenue primarily from interest rates charged to credit card holders who fail to pay their loan back at the end of the month (Massoud, Saunders and Scholnick, 2011). The importance of interest rates to a bank's credit card operation is highlighted in the way they calculate their credit card divisions return on assets, diving yearly credit card profit by the average outstanding credit card balance (Ausubel, 1997).

This dependence on interest rates has resulted in a competitive environment where the profit margin on acquiring a new customer is large enough to provide banks with incentives to accept risky customers who are more likely to fail to pay their bills on time (Ausubel, 1997). However, as banks accept risker customers in the search of charging higher interest rates, the chance of said customers defaulting on their loan and subsequently declaring bankruptcy increases. This could have the possible effect of decreasing a bank’s net revenue and therefore, measures must be put in place to determine the creditworthiness of the candidate (Altman and Saunders, 1997).

Historically, decisions surrounding loans, or the assessment of credit risk has been reliant on human discretion and expert knowledge in the form of bankers (Altman and Saunders, 1997). However, due to the considerable number of decisions and the multitude of factors that are involved in the credit card issuing process human discretion alone has proven to be ineffective (Khandani et al. 2010). Instead, the financial industry has had to adapt and transfer their reliance onto algorithms and models to generate numerical scores of creditworthiness of applicants (Khandani et al. 2010). This is due to the subjective nature of human knowledge as opposed to the objective nature of algorithmic decision making.

5. Analysis approach

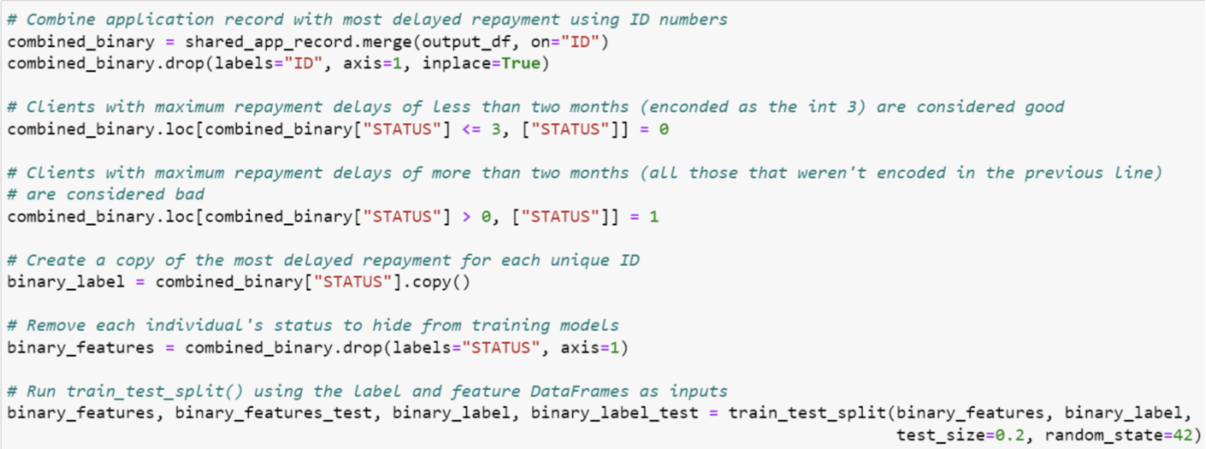
5.1 Binary Classification Problem

The first step in addressing the business question is to train a binary classification model to distinguish between "good" applicants and "bad" applicants. For the current project, good applicants were classified as individuals who had never been more than two months late on loan repayments. This cut-off point was determined based on two main factors. Initially, the cut-off for good applicants was planned to be at three months, as consumer credit default is commonly defined as delinquency beyond a period of 90 days. However, based on the observed distribution of the data, the two-month cut-off was selected due to the small number of defaults present in the database according to the three-month definition cited above.

5.1.1 Test-Train Split

In order to train a model, the available data must be split into testing and training data. In the current example, the relative sizes of these sets were determined to be 80% and 20% of the available data, respectively.

Prior to performing the split, the application record data must be merged with the credit record data into a single data frame. Following the merge, the "ID" column is dropped from the data frame, as it holds no relevant financial information. Next, the "STATUS" column is edited to reflect the binary distinction between good and bad applicants as outlined above. The now binary "STATUS" column is copied into the binary_label variable and dropped from the binary_features variable.



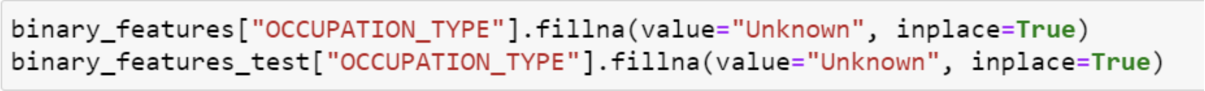
```
# Combine application record with most delayed repayment using ID numbers
combined_binary = shared_app_record.merge(output_df, on="ID")
combined_binary.drop(labels="ID", axis=1, inplace=True)

# Clients with maximum repayment delays of less than two months (enconded as the int 3) are considered good
combined_binary.loc[combined_binary["STATUS"] <= 3, ["STATUS"]] = 0

# Clients with maximum repayment delays of more than two months (all those that weren't encoded in the previous line)
# are considered bad
combined_binary.loc[combined_binary["STATUS"] > 0, ["STATUS"]] = 1

# Create a copy of the most delayed repayment for each unique ID
binary_label = combined_binary["STATUS"].copy()

# Remove each individual's status to hide from training models
binary_features = combined_binary.drop(labels="STATUS", axis=1)

# Run train_test_split() using the label and feature DataFrames as inputs
binary_features, binary_features_test, binary_label, binary_label_test = train_test_split(binary_features, binary_label,
                                                                                          test_size=0.2, random_state=42)
```

5.1.2 Data Pre-Processing

Following the test-train split, certain subsequent steps must be taken on the training data in order to ensure said data is fit to train a model. These steps can include imputing missing data points, encoding categorical features, and over- or under-sampling the training data.

Imputing -

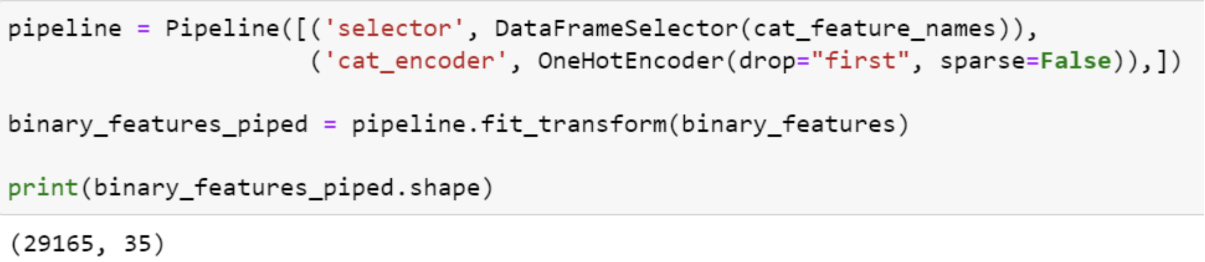
The first step is to impute any missing data. In this case, only "OCCUPATION_TYPE" has any null values. These null values were replaced with "Unknown", as it was determined that an individual's occupation could potentially be an important consideration in certain cases, so it would be inappropriate to drop the feature completely.



```
binary_features["OCCUPATION_TYPE"].fillna(value="Unknown", inplace=True)
binary_features_test["OCCUPATION_TYPE"].fillna(value="Unknown", inplace=True)
```

Encoding -

Following imputation, categorical data must be encoded numerically. For certain features, such as the gender of an individual or whether they own property, this task is trivial. However, for other categorical features, a more complex encoding method must be used. In this case, sklearn's Pipeline function will be used with One-Hot-Encoder.

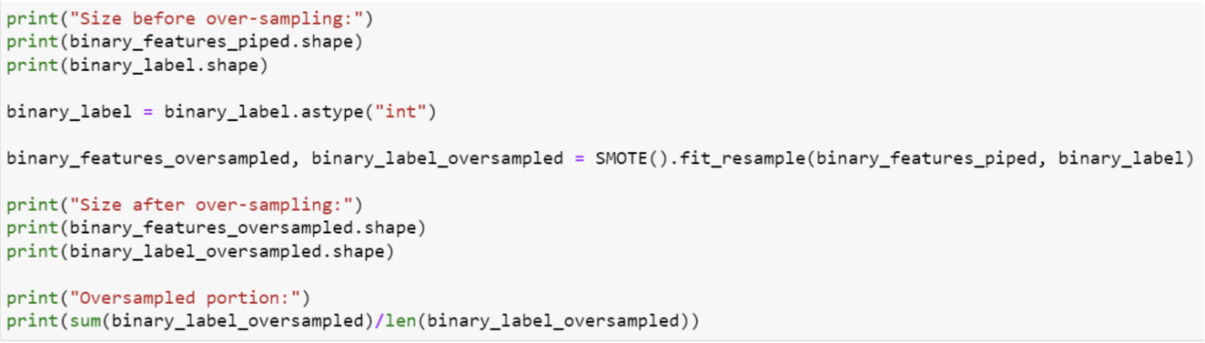


```
pipeline = Pipeline([('selector', DataFrameSelector(cat_feature_names)),
                     ('cat_encoder', OneHotEncoder(drop="first", sparse=False)),])

binary_features_piped = pipeline.fit_transform(binary_features)

print(binary_features_piped.shape)
```

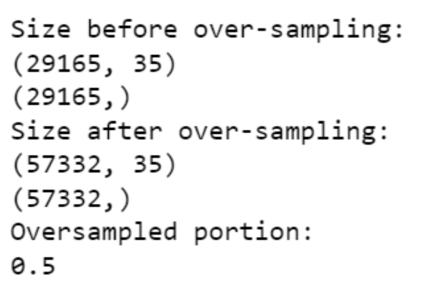
```
(29165, 35)
```

Over-Sampling -

Due to the uneven distribution of data, the training data must over-sample minority class data. To do this, random entries which fall into the “bad applicant” category will be made so that the distribution of the training data is sufficient to effectively train the models. In this case, the SMOTE().fit_resample() function was used. SMOTE was used over random over-sampling to minimize the chances the resultant models will overfit the training data.



```
print("Size before over-sampling:")
print(binary_features_piped.shape)
print(binary_label.shape)

binary_label = binary_label.astype("int")

binary_features_oversampled, binary_label_oversampled = SMOTE().fit_resample(binary_features_piped, binary_label)

print("Size after over-sampling:")
print(binary_features_oversampled.shape)
print(binary_label_oversampled.shape)

print("Oversampled portion:")
print(sum(binary_label_oversampled)/len(binary_label_oversampled))
```

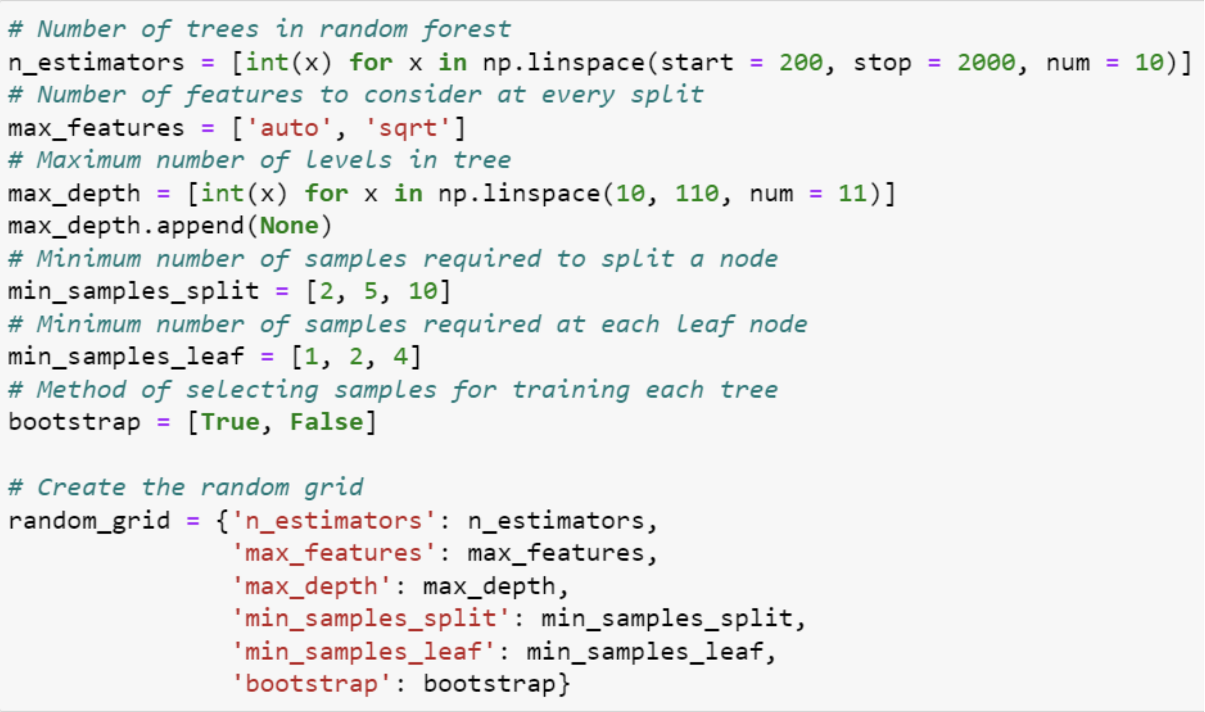


```
Size before over-sampling:
(29165, 35)
(29165,)
Size after over-sampling:
(57332, 35)
(57332,)
Oversampled portion:
0.5
```

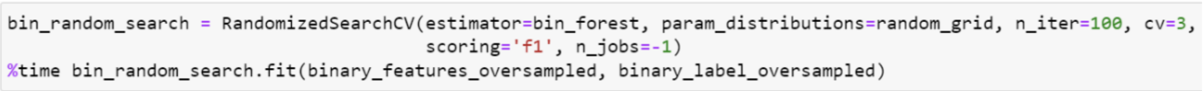
5.1.3 Training Models

Following the data pre-processing, three models were fitted to the binary testing and training data. These models are logistic regression, random forests, and support vector machines (SVM). These models are defined in section 2 above.

After training the three models, their results were compared. The random forest produced the most accurate results, but the logistic regression's interpretability was considered significant enough to warrant further investigation. The random forest model subsequently underwent hyper-parameter optimisation.

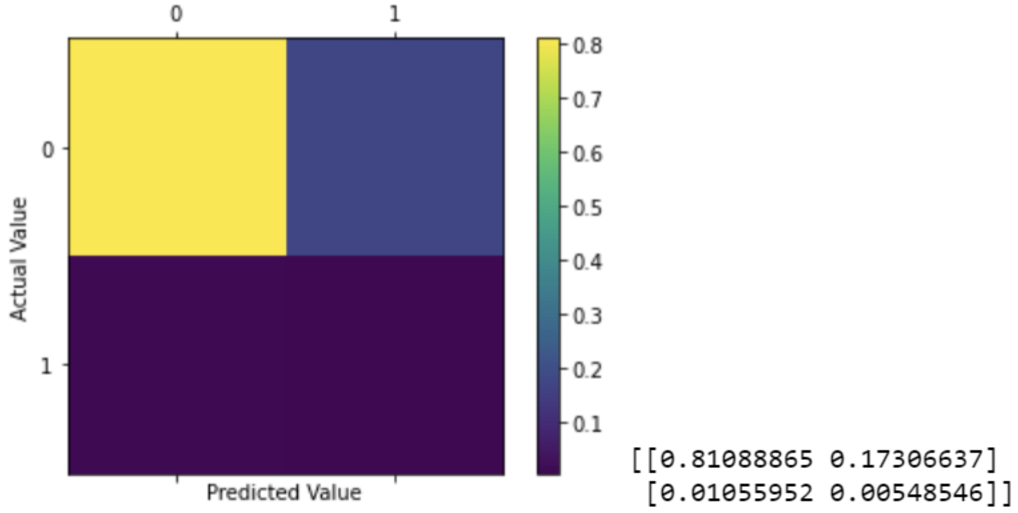


```
# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
# Method of selecting samples for training each tree
bootstrap = [True, False]

# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
```



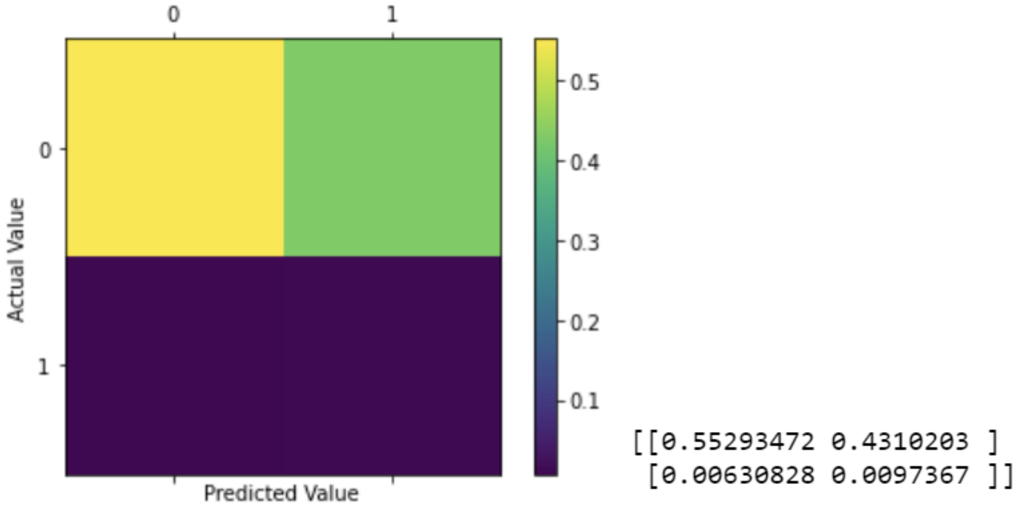
```
bin_random_search = RandomizedSearchCV(estimator=bin_forest, param_distributions=random_grid, n_iter=100, cv=3,
                                       scoring='f1', n_jobs=-1)
%time bin_random_search.fit(binary_features_oversampled, binary_label_oversampled)
```

5.1.4 Evaluation



```
[[0.81088865 0.17306637]
 [0.01055952 0.00548546]]
```

Figure 2: Random Forest Confusion Matrix



```
[[0.55293472 0.4310203 ]
 [0.00630828 0.0097367 ]]
```

Figure 3: Logistic Regression Confusion Matrix

From the confusion matrices seen above, the random forest model is much more effective at predicting true negatives (or good applicants) than the logistic regression. Additionally, the overall accuracy of the random forest is significantly higher (82% compared to 56%). However, both models perform poorly when predicting "bad customers". For both the random forest and the logistic regression, the rate of false positives is high, while the rate of true positives is low. Although these results could, in part, be due to the unbalanced nature of the original dataset, they are still less than optimal.

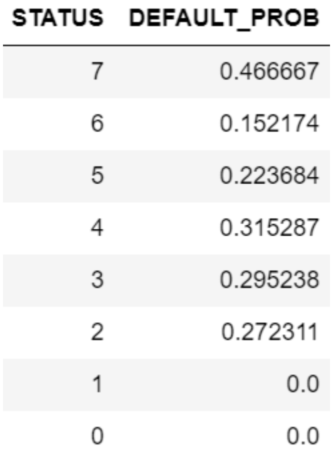
Potential reasons for this could be that the features are less correlated to the label than initially thought or are not detailed enough to provide sufficient information to accurately predict the label. For example, an applicant might own a car, but the value of the car might be more significant in determining the probability of default.

5.2 Multiclassification Problem

Following the binary problem, a multiclassification model must be designed that assesses how late an applicant will eventually pay their loans back, if at all. This model would be used to guide the decision on what interest rates to assign clients such that the repayments cover the defaults of other clients. In practice, the multiclassification model would be used after the binary model with the data from the good applicants.

5.2.1 Statistical Manipulation

To approach this problem, the statistical analysis already performed must be built upon to determine the probability that a delayed repayment will eventually result in defaults or full repayments. To do this, the data for the latest each client has repaid their loan must be taken and compared to the most recent loan data. Any case where the most recent loan data does not reflect a full repayment, the client will contribute to their latest repayment category's probability of default.



STATUS	DEFAULT_PROB
7	0.466667
6	0.152174
5	0.223684
4	0.315287
3	0.295238
2	0.272311
1	0.0
0	0.0

Figure 4: Default Probabilities for Encoded Status Categories

5.2.2 Test-Train Split and Data Pre-Processing

The test-train split and data pre-processing for the multiclassification problem were handled similarly to the binary classification problem. One notable difference, however, is that the oversampled dataset for the multiclassification problem is significantly larger due to the equal representation of the eight status categories instead of just two.

5.2.3 Training Models

The multiclassification problem also utilizes both a random forest and SVM, as these models are well suited to multiclassification distinction. Additionally, an artificial neural network was trained. These models are made up of several layers of interconnected nodes with associated weights and thresholds. These determine whether a node will output a signal and how much the signal impacts the model's final decision. Artificial neural networks can be among the most accurate machine learning models, particularly for complex tasks like computer vision and language processing, but they are difficult to interpret and thus explain.

After training and comparing these three models, it was determined that the random forest produced the best results. As such, it was optimized using the same algorithm as was used in the binary classification problem.

The SVM was significantly less accurate than the random forest and took far too long to train to be feasible for further evaluation. Similarly, the artificial neural network was found to have very low validation accuracy and was not considered further.

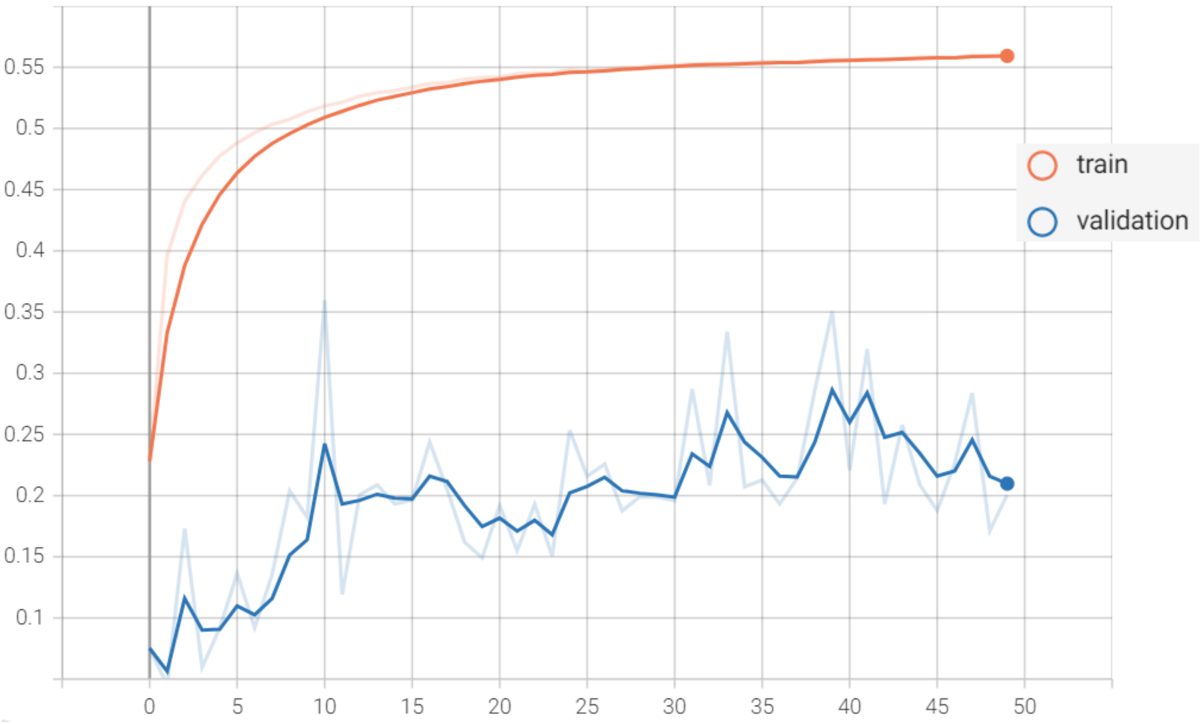


Figure 5: Artificial Neural Network Accuracy by Epoch

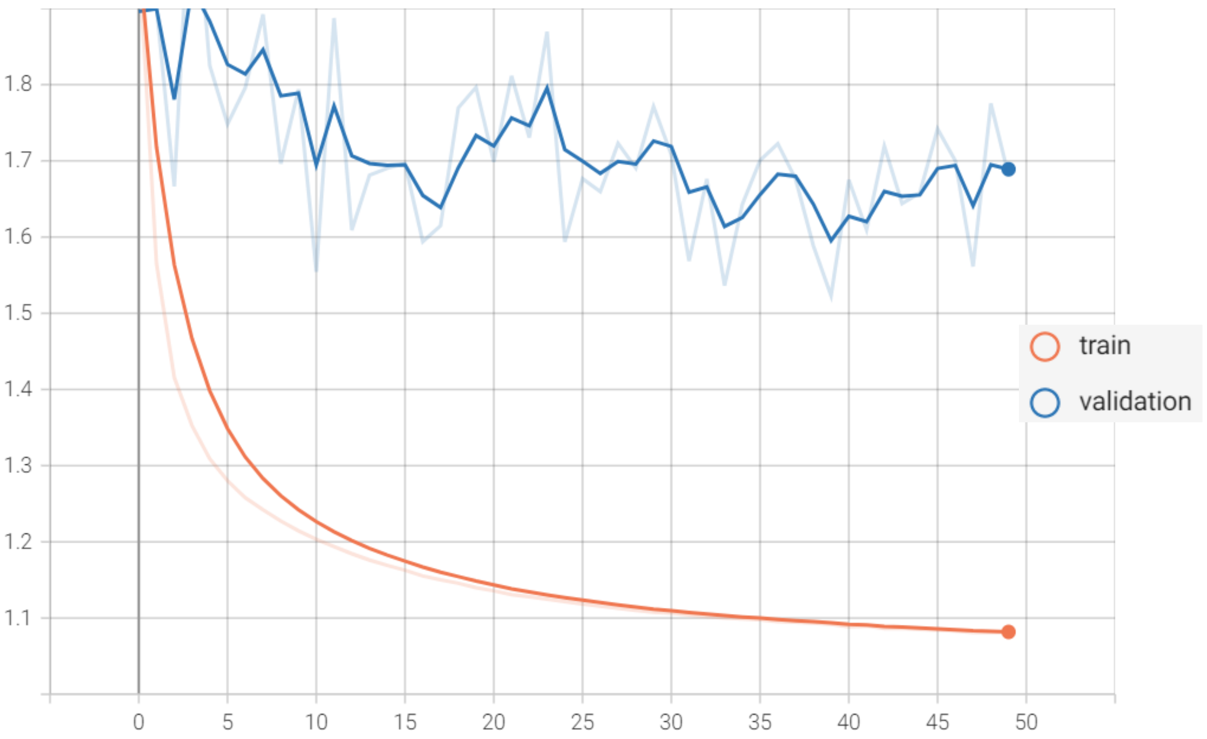
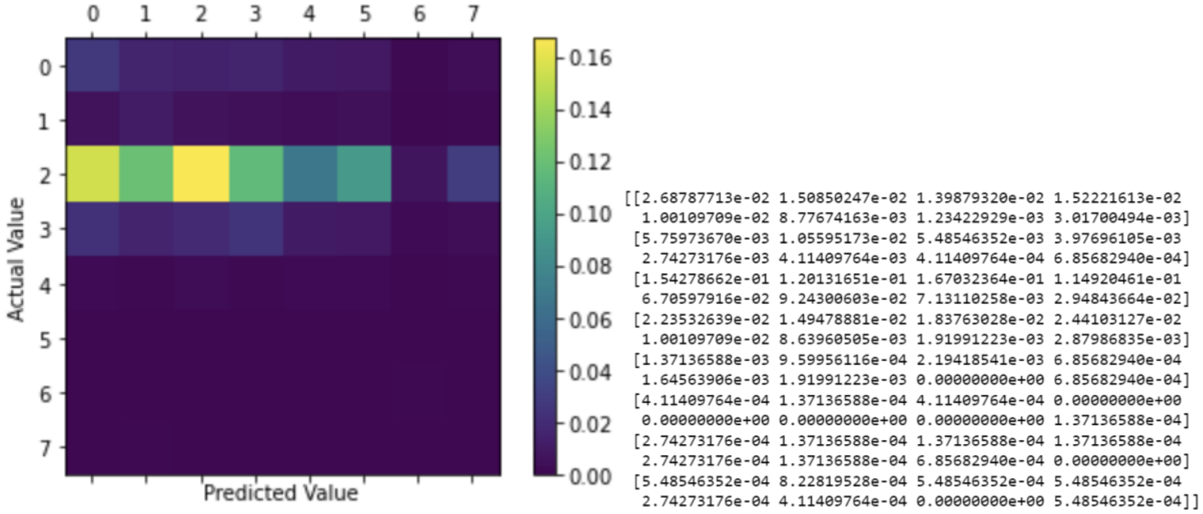


Figure 6: Artificial Neural Network Loss by Epoch

5.2.4 Evaluation



```
[[2.68787713e-02 1.50850247e-02 1.39879320e-02 1.52221613e-02
  1.00109709e-02 8.77674163e-03 1.23422929e-03 3.01700494e-03]
 [5.75973670e-03 1.05595173e-02 5.48546352e-03 3.97696105e-03
  2.74273176e-03 4.11409764e-03 4.11409764e-04 6.85682940e-04]
 [1.54278662e-01 1.20131651e-01 1.67032364e-01 1.14920461e-01
  6.70597916e-02 9.24300603e-02 7.13110258e-03 2.94843664e-02]
 [2.23532639e-02 1.49478881e-02 1.83763028e-02 2.44103127e-02
  1.00109709e-02 8.63960505e-03 1.91991223e-03 2.87986835e-03]
 [1.37136588e-03 9.59956116e-04 2.19418541e-03 6.85682940e-04
  1.64563906e-03 1.91991223e-03 0.00000000e+00 6.85682940e-04]
 [4.11409764e-04 1.37136588e-04 4.11409764e-04 0.00000000e+00
  0.00000000e+00 0.00000000e+00 0.00000000e+00 1.37136588e-04]
 [2.74273176e-04 1.37136588e-04 1.37136588e-04 1.37136588e-04
  2.74273176e-04 1.37136588e-04 6.85682940e-04 0.00000000e+00]
 [5.48546352e-04 8.22819528e-04 5.48546352e-04 5.48546352e-04
  2.74273176e-04 4.11409764e-04 0.00000000e+00 5.48546352e-04]]
```

Figure 7: Multiclass Random Forest Confusion Matrix

As can be seen by the confusion matrix above, the random forest was not very accurate in its predictions of the testing data. This seems to be a case of overfitting, as the final random forest model was found to only be 23% accurate on the testing data, compared to 62% on the training data.

Potential solutions to these less-than-optimal results are like those provided for the binary problem. More numerical data would perhaps lead to better results than the categorical flags present in the current dataset. In the case of an actual bank issuing credit cards, it is likely this numerical data would already be available.

6. Business Implications

Using pre-trained models for candidate selection has several business implications. The models trained for the purposes of this analysis will predict whether an applicant will get a loan or not. This is done by determining whether someone is a good or bad applicant. Based on our random forest model, this prediction will be 83% accurate, however, it is important to note that this figure is inflated due to the lopsided nature of the dataset (i.e. there are more good customers than bad). It is ~50% effective at identifying bad customers. The major business implications exist within this margin of effectiveness.

The first of two outcomes are that the model predicts a good applicant who turns out to be bad, by defaulting. In this scenario, the applicant may never repay the loan, or they will repay the loan late. If they never repay the loan, the bank loses both the initial sum of the loan and the interest owed. If they repay late, the bank receives the initial sum, plus interest accumulated over the defaulting period. This shows that despite the error in prediction, the bank can still profit, however, the risk of loss is increased.

The second outcome is that the model predicts a bad applicant who turns out to be good. This outcome would be more difficult to collect data on, as if the model does not recommend the applicant, they are likely not to receive a credit card. The risk is also much lower in this scenario, as the bank neither gains nor loses. This mitigates the risk, but it does result in a loss of income.

Automation and the use of models in business decision making do have implications for the business, however, human error also exists and must be acknowledged. There is no perfect solution and misjudgements are bound to occur. Even though less than 50% of bad applicants (positives) were correctly identified, the binary random forest might still be useful. Having half the previous number of defaults is likely to be beneficial to any credit card provider. Further analysis would be needed to determine whether this decrease in defaults would be worth the potential decrease in clients who will be able to repay in full.

As a result of its poor performance, it cannot be recommended that the multiclass random forest model is applied to the business problem, at least in its current state. While it is possible that more numerical data would help increase the accuracy of the multiclassification random forest, a better approach might be to create an ensemble of One-Versus-All classifiers. In such a case, each constituent classifier would work as a binary classification model for one status category. This would combine the efficacy seen in the binary classification problem with the increase in flexibility afforded by distinguishing between more specific classes.

7. Limitations and future research

The first limitation presented by this project is the possibility of bias inherent in the dataset. This project used a secondary source dataset to train the models used. If the training data is biased, the machine learning system trained on it will learn the bias and thus replicate the bias in its outcomes (Pena et al., 2020). Additionally, this data may also include association bias, this occurs when training data multiples the cultural bias already prevalent in the organisation (Chou et al., 2017). A further limitation is the possibility of human bias, as the models are constructed by humans and rely on data created by humans (Mujtaba, 2019). This is known as interaction bias, caused by the interaction of humans and the models they create (Chou et al., 2017). Bias can be mitigated by understanding the background of your data, ensuring that the data and model creation follows policies and procedures to avoid bias and regularly testing your models and data for bias.

A further limitation is the dangers of false prediction, if a false prediction is made, the bank may lose potential revenue, or will lose both the interest payable and the initial sum of loans given. There is also the possibility of additional revenue, but this comes with greater risk. The final predominant limitation the project has identified is the lack of available features in the dataset. Given the secondary nature of the dataset, the features are predetermined. There are features missing in the dataset which the project team believe should be factored into the decision-making process. Should future research occur, the project team urge the researchers to find a more complete dataset or create a dataset with all relevant features for more accurate results and modelling.

There was no available data on the interest rates offered to credit card loan receivers. This project would have been able to make much more accurate models and predictions if data on interest rates would have been available. These would allow the models to set differing interest rates depending on the degree of risk posed to the bank by an individual applicant. Rather than saying yes or no to an applicant, the greater the risk, the higher the interest rate offered. Interest rate data, particularly individual interest rate data would significantly contribute to providing more accurate predictions and findings. Future researchers should endeavour to find interest rate data or create the data for more accurate findings.

8. Reflective Conclusion

Ultimately, the business problem dealt with in this project comes down to the detection of risk and the willingness of banks to take on this risk. As we have identified in the literature review, liberal lending strategies and an over accumulation of risk can be detrimental to a banks long term success, as evident in the 2008 financial crisis. However, as we have seen the accumulation of revenue made from late payments is a substantial source of revenue for banks. Therefore, it is important for banks to determine the level of risk that balances interest rate revenue with the possibility of defaults and the losses that come with it. We formulated the business case that we dealt with in this project, aiming to assess the likelihood of credit card loans being paid back on time, the eventually that late repayments lead to full repayments and taking industry standard credit card into account, whether to offer an applicant a credit card loan based on the expected profit that can be made.

We believe that the data set that we have presented and the accompanying data exploration, data preparation, model selection and model evaluation that followed successfully addresses the business problem at hand and helps the bank accurately identify the level of risk that optimises revenue and profit. More specifically the initial binary model that we ran, which made use of random forest models, yielded effective predictions of true negatives (good applicants) with higher overall accuracy. However, unfortunately the same cannot be said for its performance in predicting bad customers.

9. References

Altman, E.I. and Saunders, A. (1997). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance*, [online] 21(11-12), pp.1721–1742. doi:10.1016/s0378-4266(97)00036-8.

Ausubel, L.M. (1997). *Credit Card Defaults, Credit Card Profits, and Bankruptcy | ECON l Department of Economics l University of Maryland*. [online] Umd.edu. Available at: https://www.econ.umd.edu/publication/credit-card-defaults-credit-card-profits-and-bankruptcy [Accessed 4 Aug. 2022].

Breiman, L., 2001. Random Forests. In: Machine Learning, 45(1), pp. 5-32.

Chou, J., Ibars, R. and Murillo, O. (2017). In Pursuit of Inclusive AI Five Ways to Identify Bias 5 Insights for Inclusive AI 23 Acknowledgements 39. [online] Available at:https://www.microsoft.com/design/assets/inclusive/InclusiveDesign_InclusiveAI.pdf.

Gendal (2014). *Why the payment card system works the way it does – and why Bitcoin isn't going to replace it any time soon*. [online] Richard Gendal Brown. Available at: https://gendal.me/2014/07/05/why-the-payment-card-system-works-the-way-it-does-and-why-bitcoin-isnt-going-to-replace-it-any-time-soon/ [Accessed 4 Aug. 2022].

Huang, S., Cai, N., Pacheco, P.P., Narandes, S., Wang, Y. and Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics & Proteomics*, 15(1). doi:10.21873/cgp.20063.

Kaggle (2020). *Credit Card Approval Prediction*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction?select=application_record.csv [Accessed 8 Aug. 2022].

Khandani, Amir E, Kim, A.J. and Lo, A.W. (2010). Consumer Credit-Risk Models Via Machine-Learning Algorithms. *Mit.edu*. [online] doi:0378-4266.

Kim, H., Cho, H. and Ryu, D. (2018). An empirical study on credit card loan delinquency. *Economic Systems*, 42(3), pp.437–449. doi:10.1016/j.ecosys.2017.11.003.

Kühnlein, M., Appelhans, T., Thies, B. and Nauss, T. (2014). Improving the accuracy of rainfall rates from optical satellite sensors with machine learning — A random forests-based approach applied to MSG SEVIRI. *Remote Sensing of Environment*, 141, pp.129–143. doi:10.1016/j.rse.2013.10.026.

Massoud, N., Saunders, A. and Scholnick, B. (2011). The cost of being late? The case of credit card penalty fees. *Journal of Financial Stability,* [online] 7(2), pp.49–59. doi:10.1016/j.jfs.2009.12.001.

Mujtaba, D. and Mahapatra, N., 2019. Ethical considerations in AI-based recruitment.. IEEE International Symposium on Technology and Society (ISTAS), pp. 1-7.

Pena, A., Serna, I., Morales, A. and Fierrez, J. (2020). Bias in Multimodal AI: Testbed for Fair Automatic Recruitment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. pp.28–29.

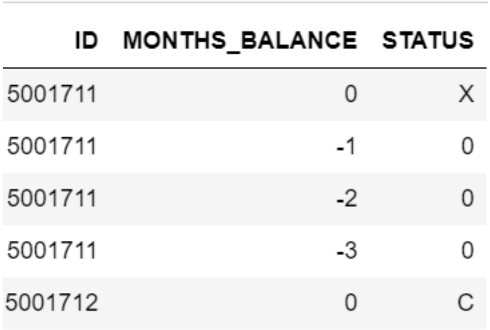
Sharma, S. and Maddulety, K. (2019). Machine Learning in Banking Risk Management: A Literature Review. *Risks*, 7(1), p.29. doi:10.3390/risks7010029.

Stoltzfus, J.C. (2011). Logistic Regression: A Brief Primer. *Academic Emergency Medicine*, 18(10), pp.1099–1104. doi:10.1111/j.1553-2712.2011.01185.x.

Zhang, Y., Lu, S., Zhou, X., Yang, M., Wu, L., Liu, B., Phillips, P. and Wang, S. (2016). Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree,k-nearest neighbors, and support vector machine. *SIMULATION*, 92(9), pp.861–871. doi:10.1177/0037549716666962

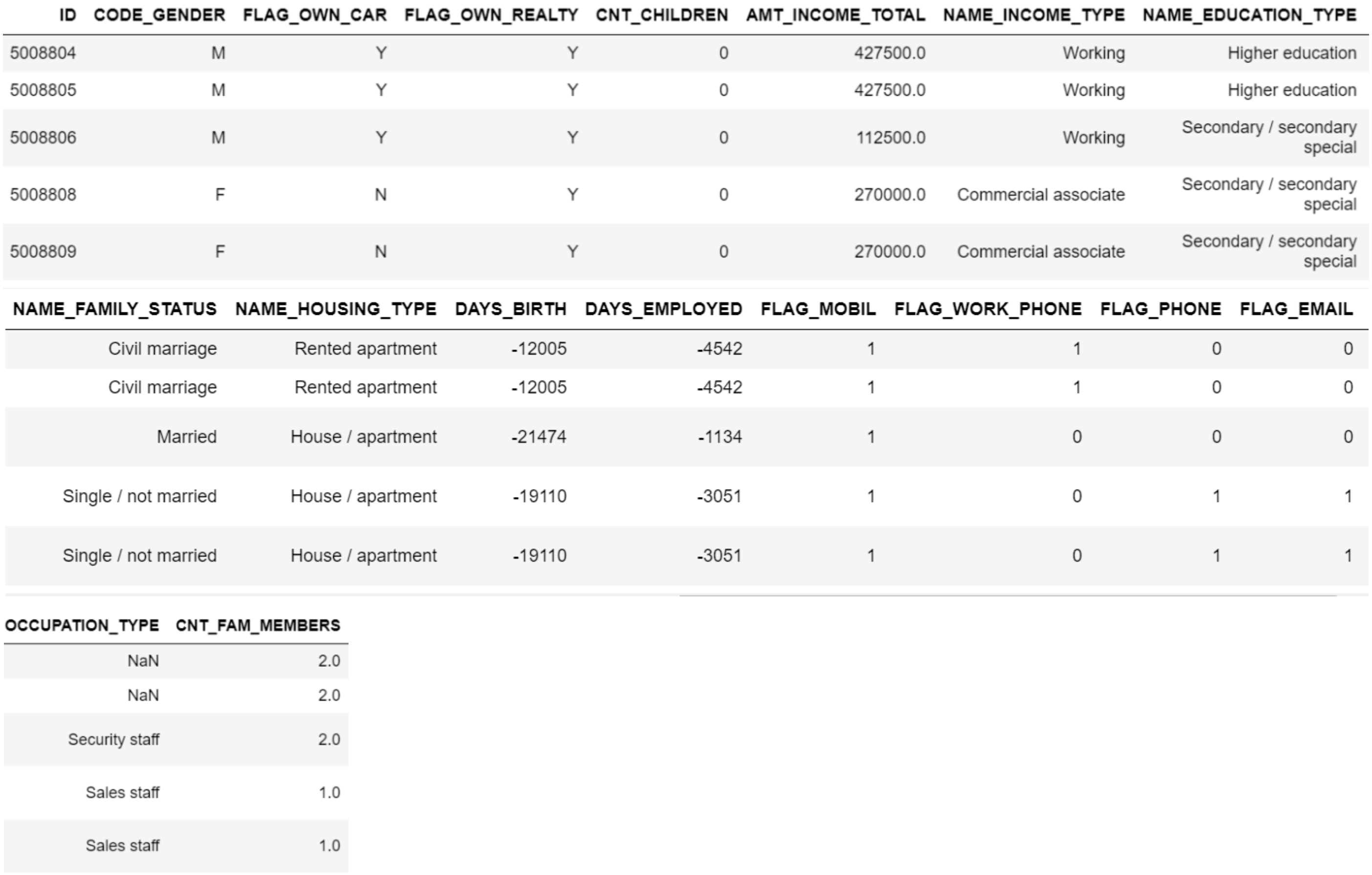
10. Appendix

Sample Credit Record Data



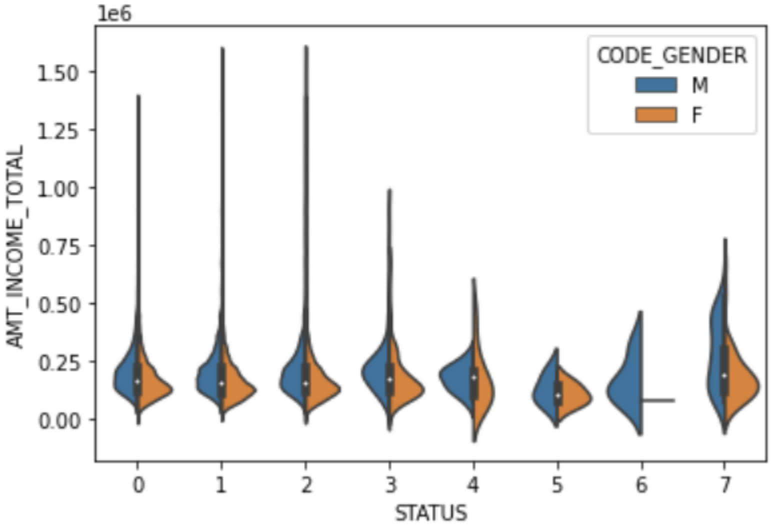
ID	MONTHS_BALANCE	STATUS
5001711	0	X
5001711	-1	0
5001711	-2	0
5001711	-3	0
5001712	0	C

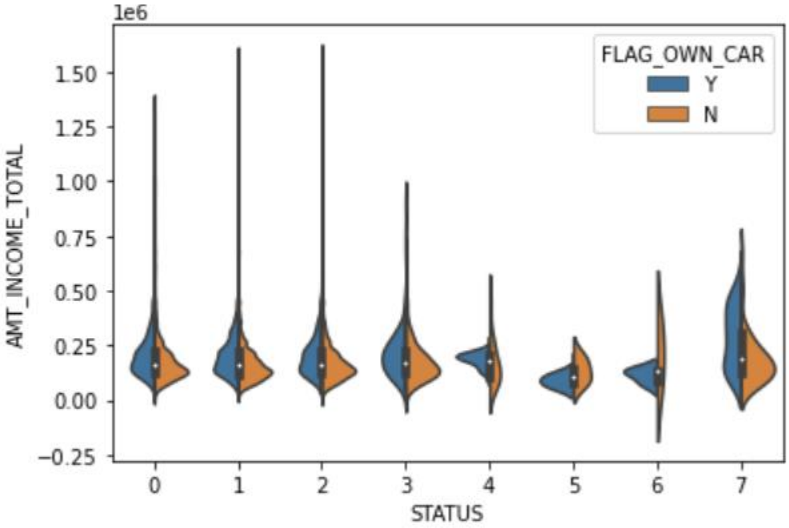
Sample Credit Feature Data

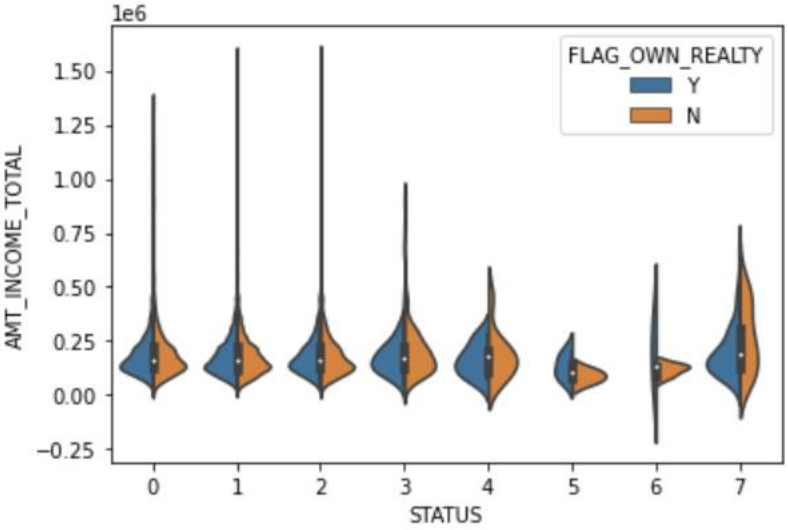


ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	DAYS_BIRTH	DAYS_EMPLOYED	FLAG_MOBIL	FLAG_WORK_PHONE	FLAG_PHONE	FLAG_EMAIL	OCCUPATION_TYPE	CNT_FAM_MEMBERS
5008804	M	Y	Y	0	427500.0	Working	Higher education	Civil marriage	Rented apartment	-12005	-4542	1	1	0	0	NaN	2.0
5008805	M	Y	Y	0	427500.0	Working	Higher education	Civil marriage	Rented apartment	-12005	-4542	1	1	0	0	NaN	2.0
5008806	M	Y	Y	0	112500.0	Working	Secondary / secondary special	Married	House / apartment	-21474	-1134	1	0	0	0	Security staff	2.0
5008808	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	Single / not married	House / apartment	-19110	-3051	1	0	1	1	Sales staff	1.0
5008809	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	Single / not married	House / apartment	-19110	-3051	1	0	1	1	Sales staff	1.0

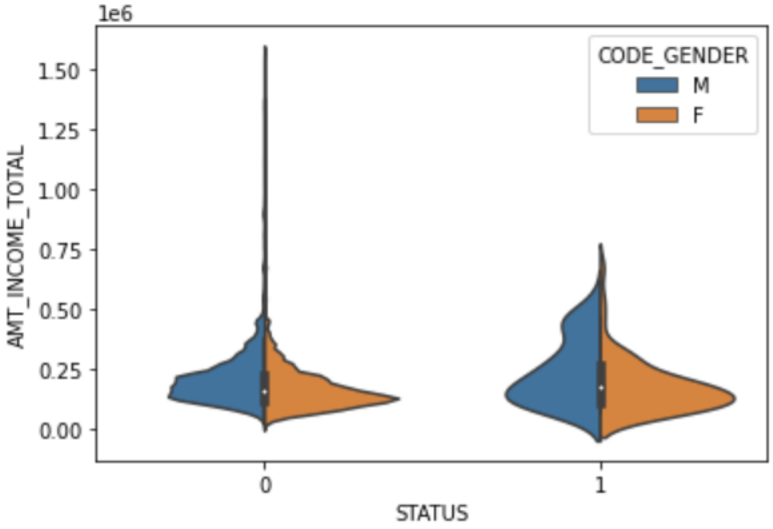
Most Recent Loan Data Violin Plots, Multiclass

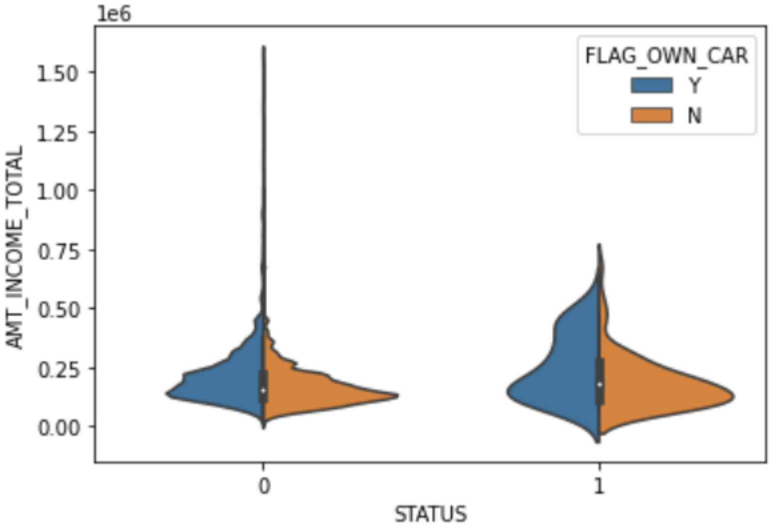


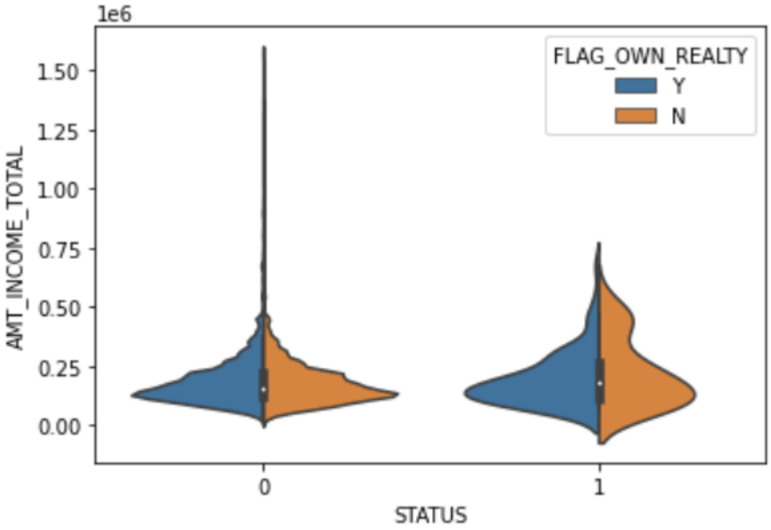




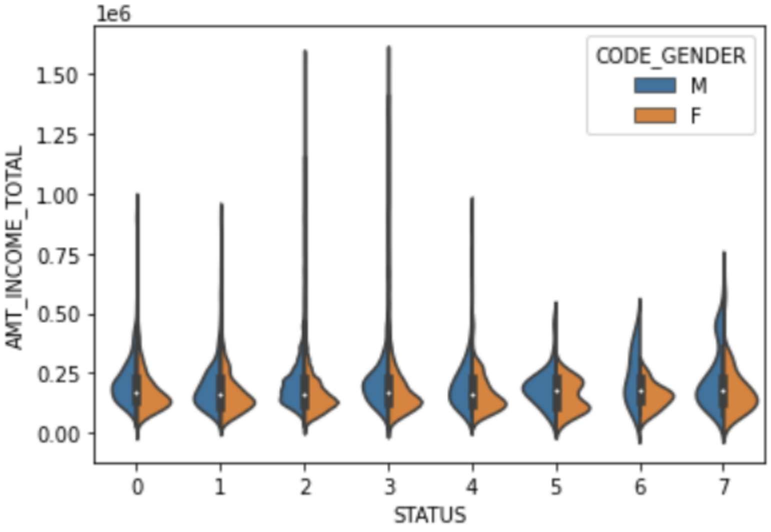
Most Recent Loan Data Violin Plots, Binary

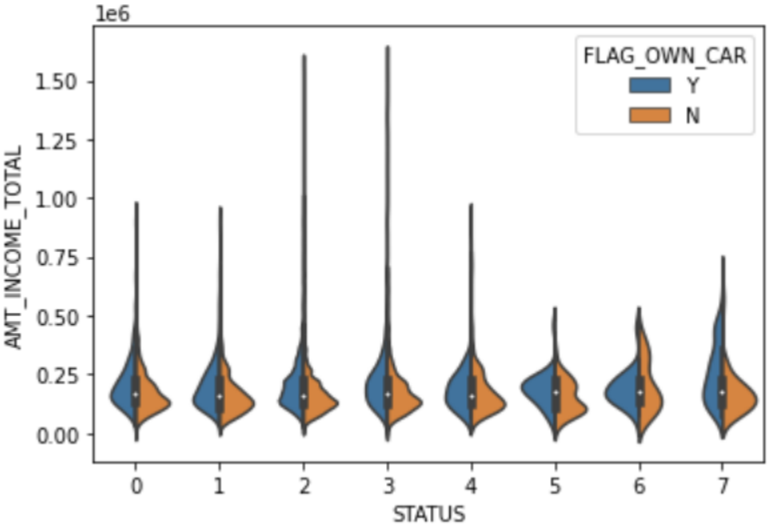


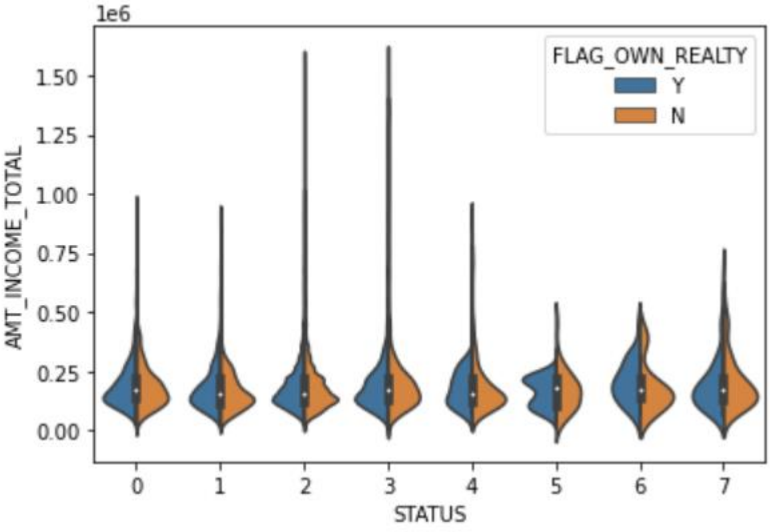




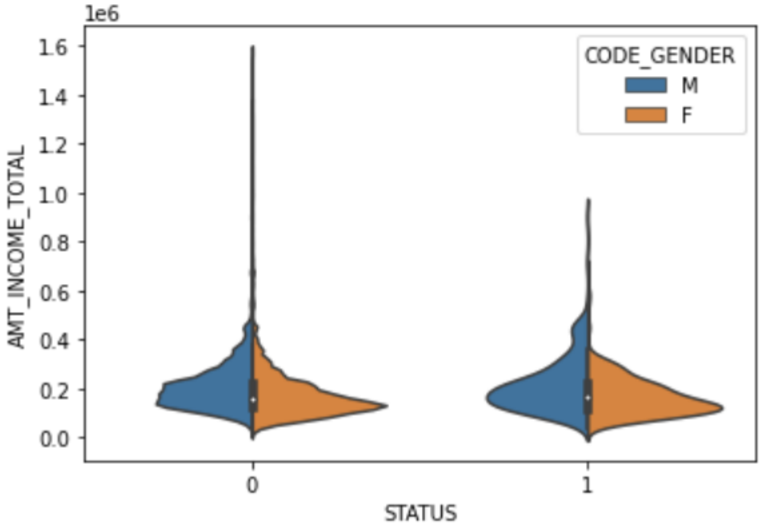
Most Delayed Loan Data Violin Plots, Multiclass

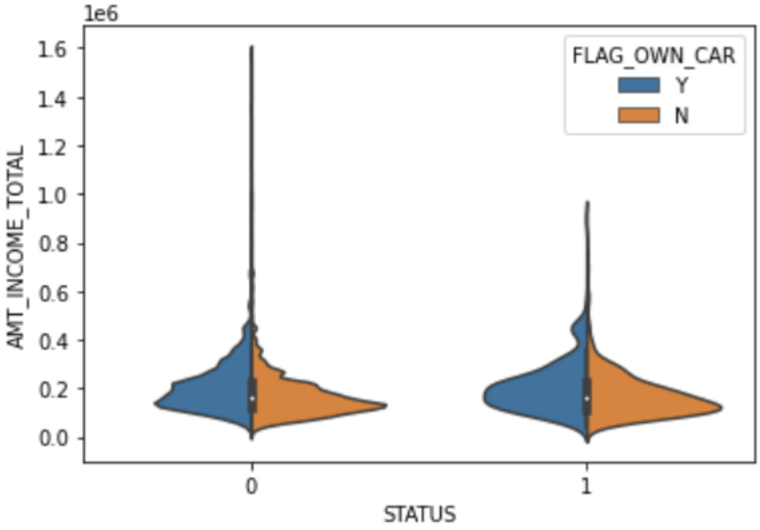


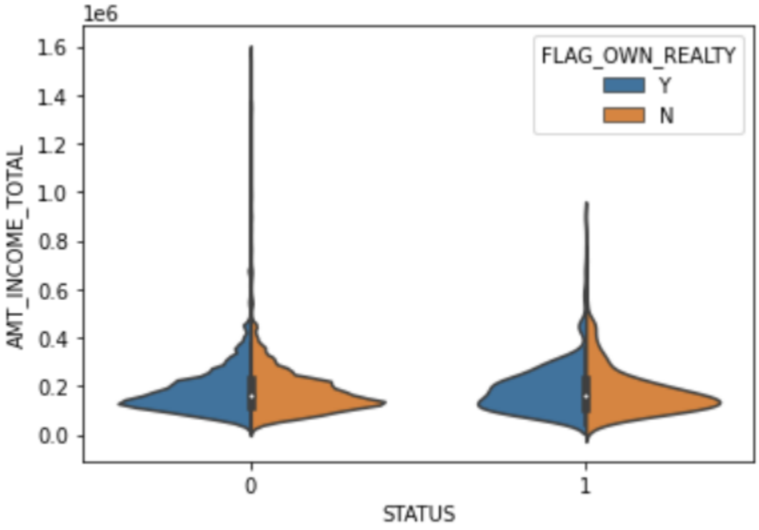




Most Delayed Loan Data Violin Plots, Multiclass







The deployed code for this project can be found at the following GitHub link:
https://github.com/nsummers98/Credit-Card-Loan-Application.git