|  |  |
| --- | --- |
|  | Machine Learning: Credit Card Information Analysis |
|  |  |
|  | Jessica Mensah - 19121729  MSc in Artificial Intelligence  1/15/24 |

**Introduction:**

With the rapid development of credit card availability, the need to critically analyse the financial risks which come with a credit card application is detrimental to a consumer and a corporation. Reducing the risk of credit card defaults inclines the need for factors such as the affordability to always be under supervision. At times, more lenient standards have allowed consumers to borrow more than they can repay. However, with stringent credit searches, companies would be able to gauge the eligibility of a consumer as responsibilities such as house ownership, the amount of dependants they have, family members as potential guarantors are vital to the amount payable a month and interest rates on loans.

Instances such as the credit card crisis in 2008 caused many house-loan defaults which is the highest responsibility of consumers monthly payments. This shows that the prediction of credit card profile analysis is an integral process before the approval of a consumer (Yu, August 2020).

**PROBLEM STATEMENT, OBJECTIVE & APPROACH**

House Ownership Prediction:

Problem Statement: Owning a house could be detrimental to the borrowing amount of a consumer. Credit worthiness of borrowers can be identified by the 5 C’s, specifically Capacity (Baiden, 2011). Capacity measures a borrower's ability to repay a loan by comparing income against recurring debts and assessing the borrower's debt-to-income (DTI) ratio. The advantage of owning a house would be reliability of an individual being able to pay debts back which would be proven by payments records. However, a disadvantage would be that a mortgage is a liability which must be paid monthly, as it is a necessity this may cause a strain on credit card payments. The first problem would be to identify how many of the applicants own their property within the dataset. This could be in the interest of credit governance whereas a credit company would like to identify if credit card applicants have any other liabilities which may lead to the need for a credit card.

Objective: A predictive analysis will be made to classify credit card applications into individuals who own a house or not and how the counter labels affect the final output.

Approach: Decision tree classification will be used to signify the programmes thought process behind the binary of an individual owning a house or not. Random Forest will be used to evaluate its own outlook regarding the correlation impact and verification of the labels which have the biggest impact on the outcome given.

Income Category:

Problem Statement: The income category of applicants is valuable information to a credit card company since it is needed to factor into Credit Risk Grading which is understanding if an applicant can pay back the amount borrowed by the company. Currently, there is no one correct system for grading loans as different banks had different risk levels (Aikman, 2017). Using Machine Learning to solve such a problem would create a centralised approach and introduce deciding factors such as how much they could borrow to an individual within their risk propensity, how much the credit card company may receive monthly for the loan given, or the amount of APR on a loan (if flexible).

Objective: The main goal is to categorise income total into ranges for visualisation purposes.

Approach: The use of Linear Regression would be able to show a correlation between the income amount against the density whereas, there may not be many applications from individuals who are making over a range of £350,000 to £400,000. Gradient Boosting is an ensemble technique which further corrects itself which can improve the model performance and could have a positive effect on the MSE score. A visualisation of labels that have caused an impact to the regression model would give a clearer outlook of the outcome.

Customer Segmentation:

Problem Statement: Customer Segmentation would be able to visualise a wider look into what type of target audience they have and their behaviours to then have a better fixation on marketing ads to their customers. Clustering customers into a Segmental Grading System can measure credit risk and differentiate groups of credits by the risk they pose. This allows bank management and examiners to monitor changes and trends in risk levels (Habtamu, D. 2019).

Objective: To randomly group the data based on their values and not their labels to give an accurate grouping system of the given categories.

Approach: The use of K- means clustering would identify natural groupings or segments of credit card applicants. In comparison, the use of Hierarchical Clustering would be able to present the relationships in a descending position.

**Dataset Description:**

The dataset was downloaded as a csv file named as ‘application record’. The dataset consists of information about the applicant, such as the gender, date of birth and education type. There are 18 variables split into 12 categorical variables, 5 continuous variables and 1 ‘ID’ variable.

|  |  |
| --- | --- |
| **Feature Name** | **Explanation** |
| ID | Unique Key |
| CODE\_GENDER | If the individual is a man or woman |
| FLAG\_OWN\_CAR | If the individual owns a car |
| FLAG\_OWN\_REALTY | If the individual owns a house/real estate |
| CNT\_CHILDREN | The number of children the individual has |
| AMT\_INCOME\_TOTAL | The amount the individual earns |
| NAME\_INCOME\_TYPE | The type of work they do for an income |
| NAME\_EDUCATION\_TYPE | The highest form of education they have partaken in |
| NAME\_FAMILY\_STATUS | If the individual is married, single, a widow, or in a civil partnership |
| NAME\_HOUSING\_TYPE |  |
| DAYS\_BIRTH | The age of the individual since the creation of the dataset in days |
| DAYS\_EMPLOYED | The number of days the individual has been employed |
| FLAG\_MOBIL | If the individual has a mobile number |
| FLAG\_WORK\_PHONE | If the individual has a work phone |
| FLAG\_PHONE | If the individual has a phone number |
| FLAG\_EMAIL | If the individual has an email address |
| OCCUPATION\_TYPE | What the individual’s occupation is e.g. student, pensioner, full time, part time |
| CNT\_FAM\_MEMBERS | The number of family members the person has |

**Data exploration**

Data exploration is used to uncover patterns and anomalies in the dataset.

A screenshot of a computer

Description automatically generated

Figure shows the data after cleaning and Label Encoding

As shown, Figure 1 represents the refined, full dataset. The average and standard deviation of the variables can provide the varying outlook of the labels. The average represents the typical value in each label and was calculated using ‘avg’ in pandas. The standard deviation was calculated using ‘std’ in pandas to measure the dispersion of the data. This can highly portray the relevance or reliability of the label if the standard deviation is high. The other labels weren’t included as they were in binary figures so wouldn’t be relevant.

* **CNT\_CHILDREN:** On average, there are 0.4 children for every individual in the dataset, this shows that many of the individuals that are applying for a credit card do not have the responsibility of children. This affects the threshold that they may be able to receive from a credit card company as they have other liabilities to consider when more capital is acquired. The standard deviation for this label was 0.65 of children meaning in the dataset there isn’t a wide range of individuals with children as a responsibility.
* **AMT\_INCOME\_TOTAL:** The average income in the dataset was £176,850 which would affect the amount that an individual would be able to borrow, however a company would take a specialised approach by breaking down the amount total into ranges whereas a person who is earning a certain amount can borrow a specific amount. The standard deviation of for the amount total is 11.72 meaning that the disparity between pay grades across the data is quite high as individuals are earning from £0-400,000.
* **DAYS\_BIRTH:** The average age in the data set is 44 years which portrays that most of the individuals are above the age of financial responsibility which would affect their credit acceptability. The company could also find a good rapport of their previous financial transaction which would then affect the amount they would be allowed to borrow. The standard deviation was 11.7 showing there is a wide array of ages in the dataset.
* **DAYS\_EMPLOYED:** The average amount of time individuals had been employed in the dataset was 5 years and the standard deviation was 4.5 so the distribution of employment is at a low rate due to the eligibility of working being 18 and the pension age being 65 years old.
* **CNT\_FAM\_MEMBERS:** The average amount of family members was 2 per individual which may be in correlation with the number of individuals with a partner and no children or single parent households with the responsibility of one child.

**Data Cleaning & Pre-processing**

The process of data cleaning is essential to a dataset as the data can be properly manipulated into visualised graphs and used without discrepancies.

A screenshot of a computer

Description automatically generated

Figure shows the state of the data before the cleaning process.

As shown in Figure 2, the machine would not be able to easily interpret the data into all types of algorithms. The data will be converted using Label Encoder, features from the NumPy library such as ‘*np. where’,* and *‘round*’ in pandas. The insignificant parameters will also be dropped concerning the relevance of the label.

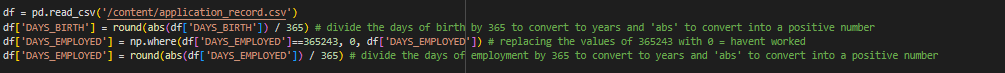


Figure shows the code used to complete the data change from negative days to positive years.

There was a need to check if there were any null values in Figure 4 which showed that the income type label had missing ‘NaN’ values which were then replaced with ‘missing’ so that it would be filled and used for some of the algorithms.

A screenshot of a computer

Description automatically generated

Figure - null values

**Label Encoder**

As many of the parameters were in the form of a string, not many algorithms can be formed with such a format. The next approach was to change the data from strings to integers with the Label Encoder command. A key was also created to interpret the graphs when created.

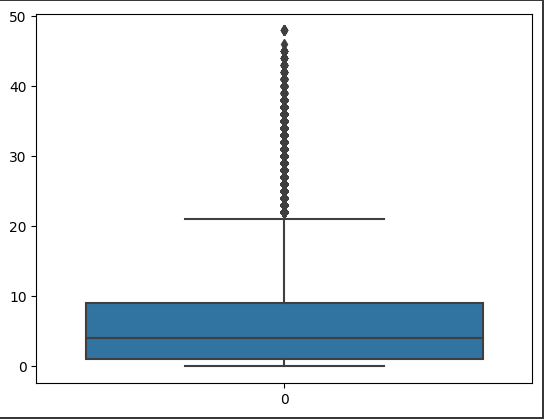
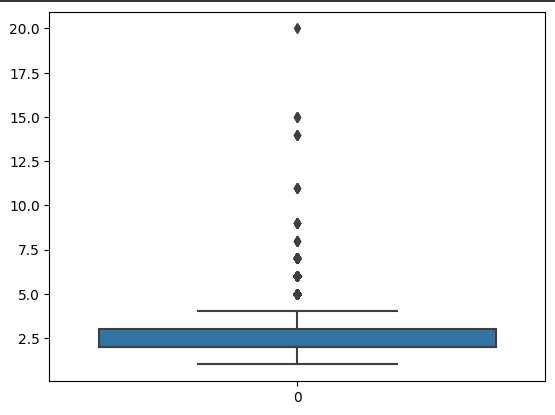
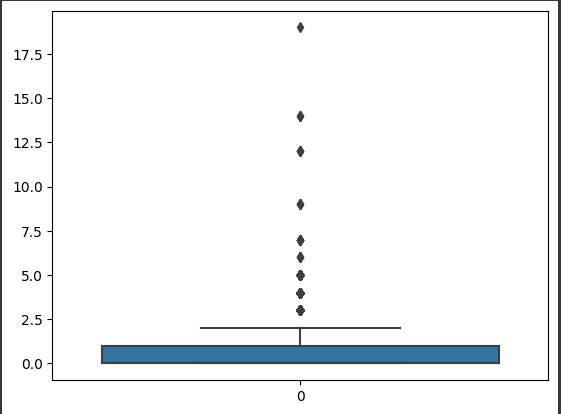
A screenshot of a computer

Description automatically generated

Figure shows the state of the data after the Label Encoding, np. where and abs.

**Outliers**

As there were many values in the data which would’ve caused the visualisation of data to be skewed, data had to be filtered as much as possible to increase the accuracy of the data. In Figures 1, the box plots were used to visualise the outliers in the data. As pictured, there were a lot of discrepancies in the data.



INCOME\_TOTAL

YEARS\_EMPLOYED

AMT CHILDREN

FAMILY MEMBERS

Figure shows the boxplots before data manipulation.

Parameters such as ‘CNT\_CHILDREN’, ‘AMT\_INCOME\_TOTAL’, ‘CNT\_FAM\_MEMBERS’ and ‘DAYS\_EMPLOYED’ were not in the best state for data manipulation which resulted in the use of identifying the upper limit through standard deviation and using the ‘loc’ method in pandas to remove the outliers from each label where necessary.

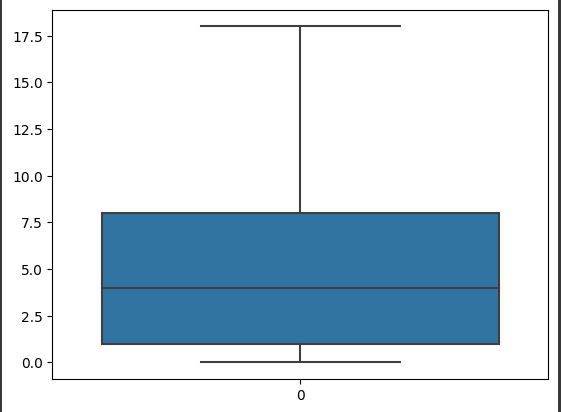
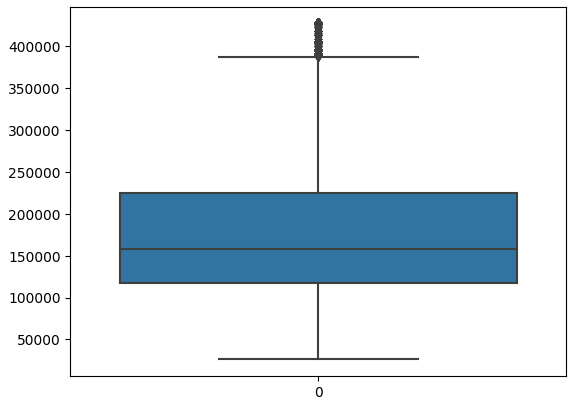
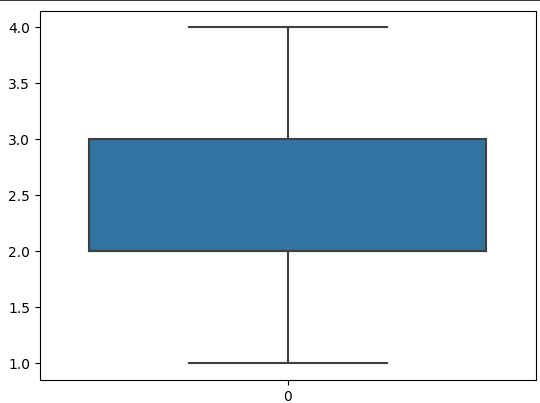
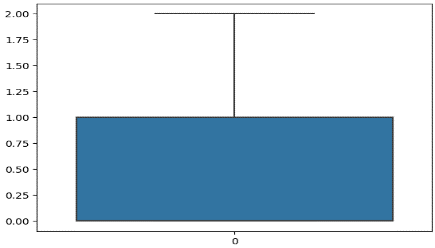
Figure - data after manipulation

INCOME\_TOTAL

DAYS\_EMPLOYED

CNT\_CHILDREN

FAMILY\_MEMBERS



**Data Visualisation**

Data Visualisation is a graphical representation of data to reveal patterns and trends. In a data-driven era, data visualisation is significant in facilitating answers and making informed decisions based on the outcome. The combination of seaborn and matplotlib was used to provide the correct visualisation graphs which can elevate exploratory data analysis by leveraging seaborn’s statistical plots and matplotlib’s plotting and ‘show’ capabilities.

**Pre-liminary data analysis**

A screenshot of a computer

Description automatically generated

Figure is a heat map regarding the full dataset.

Before any graphs were made, a full representation of the label correlation was needed to visualise which data would give an exceptional result. This two-dimensional representation of data has correlation ranges from -1 and +1, showing that values that are closer to zero would mean that there is no linear trend between the two variables. The intensity of colours represents a higher or lower correlation which allowed a selection of variables to create different visualisations.

Key:

* ‘CODE\_GENDER’: Male/Female =0/1
* ‘INCOME\_TYPE’: Working/ Commercial Associate/Pensioner/State Servant/Student = 0, 1, 2, 3, 4
* ‘EDUCATION\_TYPE’: Secondary School/ Higher Education/ Incomplete Higher/ Lower secondary/ Academic Degree = 0, 1, 2, 3, 4
* ‘NAME\_FAMILY\_STATUS’: Married/ Single/ Civil Marriage/ Separated/ Widow = 0, 1, 2, 3, 4
* ‘NAME\_HOUSING\_TYPE’: Housing/Apartment, With Parents, Municipal apartment, Rented apartment, Office apartment, Co-op apartment = 0, 1, 2, 3, 4, 5

A graph of a graph

Description automatically generated

Figure shows the income total.

From the histogram shown in Figure 7, the income of individuals varies from 50,000 to 400,000 dollars and the average income is within the bar range of $150,000 which relates to the precise average which is 176,000 dollars annually.

A graph of a number of people

Description automatically generated

Figure shows a combined kde plot of the number of days employed. Orange (1) = Female, Blue (0) = Male

As shown in figure 8, there are two factors pertaining the results which are ‘income total’ and ‘days employed (years)’. A hue of the gender was used to differentiate the data and portrays the disparity between the pay gap and the amount of time they have been employed. The ‘days\_employed’ data was converted from days to years as shown above to provide a clearer visualisation of the individuals employment longevity. Having a high number of employment years would show that an individual has a healthy flow of income which would affect the amount they may be able to acquire from a credit card company. A surprising revelation from the data which was the rate of males that were not employed was in the density of 0.4 which is the highest amount across the dataset. This may be due to life choices and factors such as being a pensioner.

A graph of blue and orange lines

Description automatically generated

Figure is a kde plot shows income total by gender.

From the KDE plot graph shown in Figure 9, there is a wide pay gap between female and male representatives in the dataset where a high density of males’ peak at £150,000, however there are bot many females who are on 150,000 and the amount peaks close to 200,000 but with a low density.

A blue pyramid shaped object

Description automatically generated

Figure shows a boxenplot of the income total.

The boxenplot in Figure 10 represents the income breakdown in the dataset and where there is a wider range of individuals within the category, the box plot is larger. As shown, many individuals in the dataset earn within the 150,000 to 250,000 range.

A diagram of different colored arrows

Description automatically generated with medium confidence

Figure shows a violin plot within family members against the family status.

A Violin plot shown in Figure 11 was used to show the density of with different family statuses with the amount of family members within their family.

A bar graph with blue and orange squares

Description automatically generated

Figure shows a bar chart showing the gender and who attains a car.

The bar chart in Figure 12 shows the amount of males or females that own a car which represents the different responsibilities and liabilities that an individual holds during the application of a credit card. The data shows that one gender may have more of a car liability than the other.

A graph showing a number of children

Description automatically generated with medium confidence

Figure - 3D Scatterplot

The multi-dimensional scatterplot shows an outlook of days employed against the number of children who are within the individual’s family, against the income total. The hue was also able to show the income type. As shown, there was no correlation between the information.

**House Ownership Prediction:**

Decision Tree – A predictive modelling algorithm which partitions data into subsets, creating a tree-like structure.

Random Forest – An ensemble method which builds multiple decision trees and outputs the mode of classes.

The actions which proceeded was to create the Decision Tree was to train, test and split the data with a test size of 30% and train size of 70%. The target variable was ‘FLAG\_OWN\_REALTY’ which is the main label against the other features in the dataset.

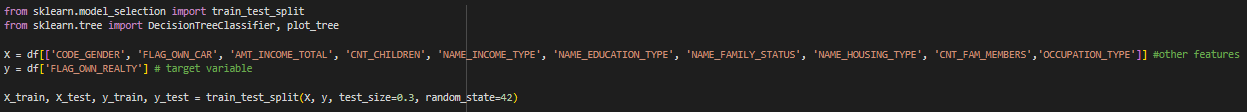


Figure - Train, test, split data

When checking the dataset, there were labelled data that did not fit the range of numbers within the dataset. For example, the Income type is Label Encoded into ‘1-20’ different incomes, however the income amount was ranging from 150,000 - 400,000 so would need to be fitted with ‘Standard Scaler’. This would allow the income amount to be added to ranges which would be labelled as a normative number.

A screen shot of a computer

Description automatically generated

Figure Normalisation of data

The use of the Decision Tree Classifier will be using the model to train the data, predict on the test and evaluate using the classification report metrics. The max depth was 6, however the amount of data shown affected the accuracy of the Decision Tree as a wider scope of information would reflect a higher score.

A screenshot of a computer program

Description automatically generated

Figure - fitting of data

it provides a quantitative measure of how well the model's predictions align with the actual outcomes in the dataset. A higher accuracy score suggests that the model is making more correct predictions and at the rate of 0.7 the model is not over fitting.

A screenshot of a computer program

Description automatically generated

Figure accuracy score

The outcome of the Decision Tree was visualised using matplotlib and shows the clear representation of the different variations of options which can come from the outcome.

A screenshot of a computer screen

Description automatically generated

Figure decision tree plot

Random Forest:

As the data has already been trained, tested, and split, there was no need to repeat the information in the next steps. Random Forest Classifier was used to further improve the accuracy of the House Ownership Prediction.

A screen shot of a computer

Description automatically generated

Figure fit and transform of model.

As shown, the accuracy score has increased by 10% with the use of Random Forest.

A screen shot of a computer

Description automatically generated

Figure accuracy of RForest

The outcome presented shows that the income of an individual highly contributes to the decision tree. Furthermore, it portrays that an individual who makes an increased amount of income can usually also own a home. The occupation type also affects the ability to own a house or not, so if an individual has a good occupation, they may have an increased permissibility to own a house.

A screenshot of a computer

Description automatically generated

Figure - barplot of RForest Feature Importance

Overall, we can depict that from the Decision tree shown, the model gives a recurring output of tree – structured outputs. Random Forest has also been able to show the feature importance which portrays the features which influence the output. To further the analysis into the Decision Tree and an understanding of its output, A bar chart was created to visualise the feature importance of the data. As shown, the ‘AMT\_INCOME\_TOTAL’ has a prominent effect on the output given in the decision tree, so most subsets in the Decision Tree would be highly influenced by this label. Another label identified is the Income Type, which inherently shows that according to the dataset, the amount of income an individual is making, including the type of income can depict if they own a piece of real estate or not.

The models will be evaluated using factors which prove the eligibility of the model’s usage.

Accuracy: Regarding the accuracy score of the models, whereas the output which closest to 1 is the model with the best accuracy score, the Decision tree model gave an output of 0.70. Random Forest gave an output accuracy score of 0.8 causing the use of Random Forest to increase the accuracy score by 0.1.

Output: The Decision Tree output would not give a visualised representation of what is occurring within the data regarding which factors influenced the output, however it would portray the most informative features for predicting the target variable. However, using the bar plot can shows an easily identifiable view of which label increases the outcome of an individual owning a property.

**Solution:** Due to the change in the accuracy score, Random Forest would give a more finite outcome.For example, a credit card company will now be able to identify that income and occupancy of an individual would greatly affect their borrowing capabilities which was identified from Machine Learning and not from biased information.

**Income Category Prediction**

A screen shot of a graph

Description automatically generatedA screen shot of a graph

Description automatically generated

Figure correlation check of labels.

As shown above, there was no correlation between the age of an individual and the income that they are making, which identifies the relevance of using machine learning models to manipulate the data into readable data. In the days employed, the scatterplot shows the same type of no correlation, however there is a mild disparity in the top right corner where the data separates. This may be due to the high amount of pay and the range of 400,000 a year is not as common.

Linear Regression – a method used to model the relationship between dependant and independent variables.

XGB Boost – Extreme Gradient Boosting leverages gradient boosting and predictive modelling tasks.

Firstly, there was a need to train, test and split the data before normalising the data as repeated in our Decision Tree with the use of Standard Scaler in sklearn preprocessing library.

A black screen with white text

Description automatically generated

Figure - test and train of data

As shown below, the data is transformed using its mean and standard deviation to standardise the train and test data. ‘X\_train\_nom’ was used to show the ranges of normalisation that each column has been converted into.

A screenshot of a computer

Description automatically generated

Figure - normalisation of data

The model was the fit and saved to files. The mean squared error score is the evaluation metric which was used and was then calculated with the testing part of the dataset. The MSE score of the regression model was very high which suggests that the model is not giving a great accuracy score of the outcome.

A screenshot of a computer program

Description automatically generatedA screen shot of a computer

Description automatically generated

Figure - fit, predict and MSE score.

A bar chart was used to identify the Feature Importance regarding the amount income label. As shown, there is no correlation between the labels and against the amount income according to the linear regression model. So, if it was represented as a scatter graph, there would be no correlation within the variables.

A screen shot of a computer

Description automatically generatedA screen shot of a graph

Description automatically generated

Figure linear regression output

The points were not correlating into any direction as the graph portrays. They are scattered far from each other which indicates that the model might not be capturing the underlying patterns in the data well.

The next step was to decrease the MSE score to receive a better accuracy score. With XGB Boost imported and fit into the model, XGB Boost was able to be used with all its features listed and ready for a new prediction.

A screenshot of a computer

Description automatically generated

Figure fitting XGBoost.

A prediction was created within the test set of data. The Mean Squared Error reduced with the use of XGB Boost meaning it was more accurate.

A screenshot of a computer program

Description automatically generated

Figure MSE score

As shown below, the feature importance was identified where the gender had an increased impact on the outcome which relates to the visualisation where males would make a high amount than females in the dataset. The education type also had an impact on the income, so this would mean, depending on the field of education, an individual could have a different salary range.

A screen shot of a computer

Description automatically generated

Figure XGBoost Output

Furthermore, regarding the ‘Values’ graph, the correlation is in a positive direction and the data becomes sparser with the increase of the value which portrays that the whilst the income decreases there are less individuals in that pay spectrum. Income ranges were presented visibly so a company would be able to see the range of customers that are in the same income category and make decisions regarding this information.

A screenshot of a computer

Description automatically generated

Figure XGBoost Linear Output

Overall, the XGB Boost can predict the values which cause an increased impact to the income total and with the combined information, a company would be able to make a balanced decision from a training model on factors such as if they are eligible for a credit card and possibly their credit limits.

* Mean Squared Error: When comparing accuracy scores, the accuracy for linear regression was **5006411980** which is not close to needed number of close to 1. However, when XGBoost was used, the accuracy decreased to **4288624396** which is a good sign that the model has recognised the use of gradient boosting, and it has made a positive impact.
* Output: With the use of Linear Regression, the machine was able to process an outcome with no real correlation between the income amount and density. However, with the use of XG Boost, the data aligned in a positive correlation.
* Feature Importance: Whilst checking for features that stood out regarding Linear Regression, it showed that all the labels stood out. However, with the use of XG Boost the feature importance was more accurate whereas the important labels were identified.
* **Solution:** The implementation of XGBoost made an immense impact where there way a visual change in the data which made it more feasible, the MSE score also reduced, so this model would be a better fit for estimating the income category of a group.

**Unsupervised problem:**

K means clustering aims to use the partitioning method that divides a dataset into 'k' distinct, non-overlapping subsets (clusters) and assigning each data point to the cluster with the nearest mean.

Hierarchical Clustering is a technique which uses a dendrogram, tree-like structure to group data points which have similarities into clusters to show the relationships between the data points.

A diagram of a diagram

Description automatically generatedA diagram of different colored dots

Description automatically generated

Figure hierarchal clustering in dendogram and scatterplot

Figure 33 shows that there is a disparity in the blue points which may show that there may be outliers in the data, however all other points are in a good position.

A screen shot of a computer

Description automatically generated

Figure silhouette (accuracy) score

The silhouette score is relatively low meaning, so the clusters were minimised which in Figure 34, decreased the silhouette score, so using 4 clusters was more accurate than a smaller amount.

A screen shot of a computer screen

Description automatically generated

Figure - 2 clusters with the new silhouette score

In contrast to the Hierarchal model, the Kmeans clustering was also performed.

The elbow method was used to identify the optimal number of clusters. This would then depict the inertia or the ‘within-cluster sum of squares’ against the number of clusters. As shown below, the datapoints would be susceptible to a range of number of clusters.

A graph with a line

Description automatically generated

Figure - elbow method used.

A screen shot of a graph

Description automatically generated

Figure K means clustering.

As shown in Figure 15, K means clustering graph the datapoint are assigned to its nearest centroid shown.

A screenshot of a computer program

Description automatically generated

Figure silhouette (accuracy) score

The silhouette score was also relatively low, however could increase depending on the number of clusters being presented.

**Accuracy Score**: Regarding Silhouette score, the closer the score is to 1, the more accurate the model is. Kmeans clustering’s accuracy score was 0.53 whereas hierarchal clustering was 0.52 so the models are close regarding accuracy.

Hierarchical clustering may be challenging to interpret visually from a presentation perspective, making its practical utility questionable. In contrast, the use of k-means clustering provides a clearer visual representation, allowing for a more straightforward understanding of which customers share similar characteristics or attributes. Due to the characteristics of the dataset and the goals of the analysis as customer segmentation, K means clustering gives a more edifying output than the Hierarchal Clustering.

**Solution:** Overall, with the right interpretation, a company would be able to visualize different customers in different segments and if the data was a recurring report, they could visualise the real-time advantage of the use of Kmeans clustering as points may move further or closer to a given centroid.

**Conclusion**

Overall, the use of Machine Learning has been able to give an insight into giving an intrinsic analysis on how different business problems can be solved, offering predictive capabilities such as house ownership, offering valuable insights such as the use of XG Boost to increase the correlation between target variables. Further development regarding hierarchical clustering would be needed to correctly label the clusters for their characteristics which would highly benefit a credit loan company. To conclude with, all the problems were solved with a functioning algorithm that would solve a business scenario.

References

Aikman, D., Kiley, M., Lee, S.J., Palumbo, M.G. and Warusawitharana, M., 2017. Mapping heat in the US financial system. Journal of Banking & Finance, 81, pp.36-64.

Baiden, J.E., 2011. The 5 Cs of Credit in the Lending Industry. Available at SSRN 1872804.

Habtamu, D., 2019. Practices of Credit Eligibility Assessment and its Perceived Relationship with Operational Efficiency: The Case of Dashen Bank SC (Doctoral dissertation, st. mary's University).

Stavins, J., 2010. Credit card borrowing, delinquency, and personal bankruptcy. New England Economic Review, pp.15-30.

Yu, Y., 2020, August. The application of machine learning algorithms in credit card default prediction. In 2020 International Conference on Computing and Data Science (CDS) (pp. 212-218). IEEE.

Appendix: https://colab.research.google.com/drive/1SGmWhVTWsD9Lcr2hS59B3LdV\_aIISGo0?usp=sharing