Portfolio Final of DM and ML

Reyna Vargas Antonio
School of Computing
National College of Ireland
Dublin, Ireland
x23127635@student.ncirl.ie

***Abstract*—Apply the knowledge acquired during the first semester of Data Mining and Machine Learning, the purpose is to analyze three different datasets, from looking for the dataset, preprocessing the information, and applying enough techniques of machine learning to develop the best model. In the first dataset, it was examined the opinion or claim of the video in the social media TikTok, whether to keep the video or not by the Logistic Regression algorithm. Following the line of opinion and claims, it was performing Naïve Baye's algorithm to analyze the reviews of the videos.**

In the second dataset, it was analyzed if the variables were correlated to suit the Multiple Linear Regression for predicting the loan amount, considering the categorical variables. Moreover, in the K-Nearest Neighbors algorithm, it was analyzed the proportion of the loan status was verified and what requirements were given more weight to get a verification as soon as possible, just like knowing what requirements help to get a loan for payments in 36 months and what is the accuracy to get it.

Implementing the Decision Tree algorithm in the third Airbnb dataset for analyzing the decision of the customer to book a listing according to the type of room (entire house/ apartment, sharing, private room, or hotel room) in San Antonio Texas.

Keywords—DM, ML, R, Multiple Linear Regression, Naïve Bayes, Logistic Regression, K-Nearest Neighbors, Decision tree, Accuracy, Confusion Matrix.

I. Introduction TikTok dataset

TikTok is a social media platform for creating, sharing, and discovering short music videos. However, TikTok is a global platform designed to give brands and marketers the solutions to be creative storytellers and meaningfully engage with the TikTok community, the big slang offers brands an opportunity for rich storytelling through a portfolio of full-screen formats that appear natively within the user experience.

TikTok was ranked No. 7 on social media platforms in 2021 and one year later was ranked No. 1. According to Search Engine Journal (Hyken, 2022), TikTok is becoming a search engine. The success of this company is to understand what the user is watching for a quick serving, offering content that is exactly what the user wants to see.

II. Introduction Loan Dataset

Loan (Reinhard, 2010) applications are merely formal loan requests (usually on a standard bank form), which are archived in the bank's records for compliance purposes. For that reason, the quantity and quality of the information directly impact a lender's credit decision. It is important to know that the basic information found on a loan application such as the borrower's name, address, phone number, employer, income, and basic property information is never enough and should always be supplemented with a complete loan request package. A loan request package is similar to a business plan and highlights the salient facts of the loan transaction.

III. Introduction AirBnB Dataset

The name Airbnb was introduced in 2009, offering grew beyond air mattresses to include spare rooms, apartments, entire houses, and more, revolutionizing the hospitality industry through the idea of matching local people with a spare room or entire home to rent to others who are visiting the area.

Airbnb serves as a community-driven platform, where exists mutual work between property owners and travelers, facilitating the rental of distinctive vacation accommodations globally. Nowadays, the company (Michelli, 2019) is focused on developing a platform, that emphasizes supporting various hosts eager to provide distinctive and memorable hospitability experiences, and it aims to connect them with an expanding community of guests who prefer personalized hospitability over standardized tourism.

IV. Related Work

The Data Mining method (Rohit Raja, 2022) helps to remove the exact significant details and develop the desired applications, it is known also as Knowledge Discovery from databases (KDD). KDD involves detecting new patterns of unknown data, and extracting the hidden pattern, data from the massive volume of the dataset, following these phases:

- Data cleaning
- Data Integration
- Data selection and Transformation
- Patter evaluation
- Knowledge representation

Similar to CRISP- DM (which is used in this project).

The application of KDD may include medical/hospitals, Marketing, Educational systems, Scientific applications, E-commerce, Retail Industries, Biological analysis, Counterterrorism, data-warehouse, in the energy sector for making, Spatial data mining, Logistics, Fraud Detection, Marketing Analysis, etc.

Machine Learning provides an alternative to algorithmic problem-solving, like giving a recipe or algorithm, taking a different approach to turning input into output. The purpose of algorithms is to examine a dataset of inputs and outputs and learn how to generate the output of its own in a process known as training.

Machine Learning means finding patterns in numbers and exploiting those patterns to make predictions, by training a model with rows or sequences to learn from the data so that, given a new sequence, the model can predict what the result

will be. Most ML models fall into one of the two board categories: supervised learning models and supervised learning models.

Supervised Learning Model, its purpose is to make predictions, training the models with labeled data so that it can take future inputs and predict what labels will be. Nevertheless, unsupervised learning models, do not require labeled data, their purpose is to provide insights into existing data, o tot group data into categories and categorize future inputs accordingly.

V. METHODOLOGY

CRISP- DM

The CRISP-DM Methodology (Delen, 2014) is made up to start with a good understanding of the business problem and the need for a data mining project and ends with the deployment of the solution that satisfies the specific business. The following steps of this methodology process:

Business Understanding: Gaining a thorough understanding of the business need for new knowledge and an explicit specification of the business objective regarding the study to be conducted.

Data Understanding: In this step is important to identify the relevant data from many available data sources.

Data Preparation: The purpose is to prepare the data for analysis using data mining methods. In general, data cleaning means filtering, aggregating, and filling in missing values.

Model Building: After already preparing the data set for modeling techniques of machine learning to address the specific business need.

Testing and Evaluation: The developed models are assessed and evaluated for accuracy, also the selected model meets the business objectives, and more models need to be developed and assessed.

Deployment: This step can be a simple report or as complex as implementing a repeatable data mining process across the enterprise. Also, include maintenance activities for the deployed models.

VI. TIK TOK DATASET

Logistic Regression

The linear regression model (David W. Hosmer Jr, 2013) is that the outcome variable in Logistic Regression is binary or dichotomous, the difference between logistic regression and linear regression is reflected both in the form of the model and its assumptions, these assumptions are:

- The response variable is binary.
- The observations are independent.
- There is NO Multicollinearity among explanatory variables.
- There are No extreme outliers.
- There is a Linear relationship between explanatory variables and the logit of the response variable.
- The sample size is sufficiently large.

A. Business Understanding

Tik Tok is a social media enterprise, so it is important to know what people are saying in the video's reviews, for that reason, by Logistic Regression it will be predicted if the video will have complaints or just normal opinions, in the way keep the video on the social media or delete it.

B. Data understanding

The TikTok dataset is a file-type CSV from the Kaggle Website, the dataset contains 19,382 observations and 12 variables. This dataset stores the columns casino, claim_status (values: claim and opinion), video_id, video_duartion, video_transciption_text (text about the video), verified_status (values: verified and not verified, author_ban_status (values: active, banned and under review), video_view_count and video_like_count. Four of the variables are qualitative data and eight of them are quantitative data. For a better understanding of the data, it will use some visualizations of them[1].

For example, figure 1, shows that few of the opinions and claims in the data have been verified, due, to still a big percentage of them, without verification. Any of the numeric variables has a normal distribution pattern like variable in the figure 2 and also, no correlation between them like in figure 3.

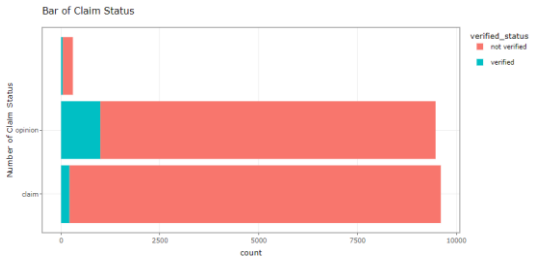


Figure 1. Bar plot of claim_status variable

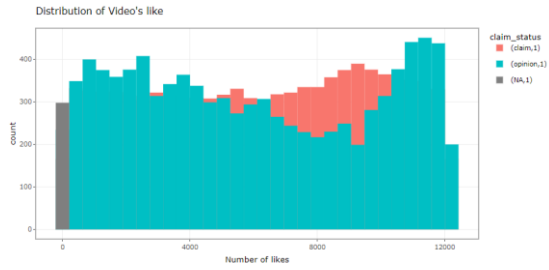


Figure 2. Histogram of Likes' Number

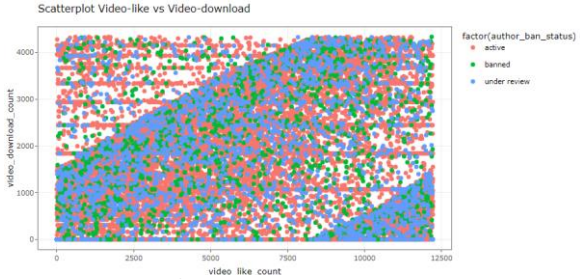


Figure 3. Scatterplot of video_like_count vs video_download_count

[1] The code file contains visualizations by categorical data, numeric data, and comparison between them.

C. *Data Preparation*

One of the most important steps to preprocessing the data, is making sure that the data is properly clean and organized so that, the first correction in this dataset is to index the variable casino as Rownames and to move the variable video_id like the first column. The next followings for checking are:

- Identify unique values in the qualitative variables for example, claim_status variables contain the values claim and opinion.
- Convert the quantitative variables into numeric data, because they were like factors.
- Handling missing values, the dataset has 596 Null values for that, the figure below do not show any percentage of missing values.

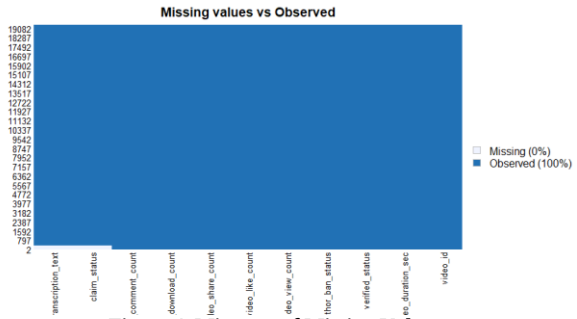


Figure 4. Missmap of Missing Values

- Fortunately, the numeric data do not have any outliers.

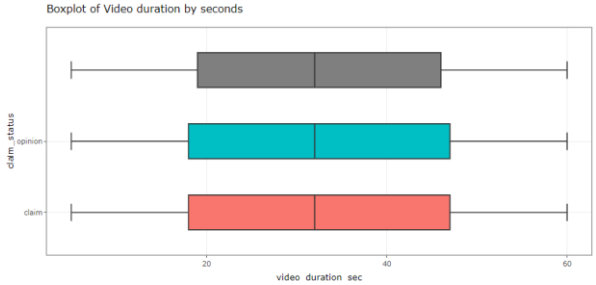


Figure 5. Boxplot of video_duration_sec (Outliers)

- There is no correlation between them as it was mentioned in the data understanding, another way to prove that is by correlogram plot.

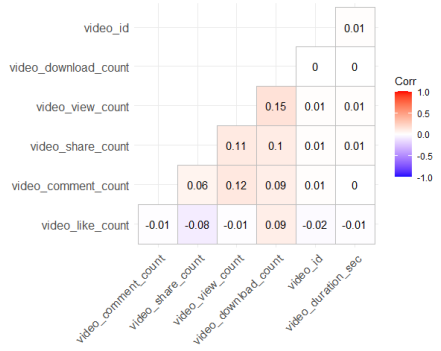


Figure 6. Correlogram

- By dummyVars, the categorical data are converted into binary variables: claim_status.opinion, verified_status.verified, author_ban_status.banned and author_ban_status.under.review.

The cleaned and transformed dataset has four columns more, which means, 15 columns and 19,382 observations.

D. *Model Building*

The dataset is split into a train dataset for training the Logistic Regression model and testing the final model in the test dataset. After, it has been verified the assumptions of the model and it considers the next steps for the building of the model:

1. The response variable is **claim_status.opinion**.
2. The explanatory variables are **video_duration_sec, video_view_count, video_like_count, video_share_count, author_ban_status.under.review**
3. The first model has obtained the variable video_share_count [2]with a P-value of 0.148, for that, the variable has been removed in the second model.

The second model is fitting better because all the explanatory variables are under the P-value (0.5).

```
Call:
glm(formula = claim_status.opinion ~ video_view_count + video_like_count +
    video_download_count + video_comment_count + verified_status.verified +
    author_ban_status.banned + author_ban_status.under.review,
    family = binomial(), data = train)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    1.810e+00  6.645e-02  27.240  < 2e-16 ***
video_view_count               2.893e-05  4.670e-06   6.194 5.85e-10 ***
video_like_count              -3.078e-05  5.667e-06  -5.432 5.58e-08 ***
video_download_count          -3.277e-04  1.626e-05 -20.154  < 2e-16 ***
video_comment_count           -1.173e-03  2.719e-05 -43.127  < 2e-16 ***
verified_status.verified       1.626e+00  1.039e-01  15.654  < 2e-16 ***
author_ban_status.banned      -2.227e+00  9.934e-02 -22.422  < 2e-16 ***
author_ban_status.under.review -1.543e+00  7.376e-02 -20.925  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18598  on 13415  degrees of freedom
Residual deviance: 14083  on 13408  degrees of freedom
  (216 observations deleted due to missingness)
AIC: 14099

Number of Fisher Scoring iterations: 5
```

Figure 7. Model 2 by Logistic Regression

E. *Test and Evaluation: Logistic Regression*

The accuracy of the model is 72.72% according to ROC Curve, which means, the model has a good predictive ability over 0.5. The Confusion Matrix confirms the accuracy of the model and displays the following information:

- True Positive (TP): 2,103 number of the correctly predicted claims (0)
- True Negative (TN): 2,013 number of the correctly predicted opinions
- False Negative (FN): 768 number of claims incorrectly predicted as opinions.
- False Positive (FP): 775 number of opinions incorrectly predicted as claims.
- Sensitivity says that 0.7325 of the proportion by the actual positive cases (opinions) that were correctly identified by the model.
- Specificity says that 72.20% of actual negative cases (claims) were correctly identified by the model.

[2] For more information about the first model, review the coding file in R.

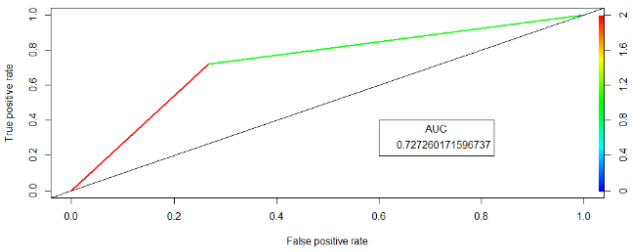


Figure 8. ROC Curve

```
predic    0    1
     0 2103  775
     1  768 2013

               Accuracy : 0.7273
                 95% CI : (0.7155, 0.7389)
    No Information Rate : 0.5073
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4545

 Mcnemar's Test P-Value : 0.8786

            Sensitivity : 0.7325
            Specificity : 0.7220
         Pos Pred Value : 0.7307
         Neg Pred Value : 0.7238
             Prevalence : 0.5073
         Detection Rate : 0.3716
   Detection Prevalence : 0.5086
      Balanced Accuracy : 0.7273

       'Positive' Class : 0
```

Figure 9. Confusion Matrix

Naïve Bayes Classifier

The Naïve Bayes Classifier (Gary D. Miner, 2012) makes the very simplistic assumption that all of the objects to be classified are completely independent of one another in the terms used to characterize them, this classifier can be remarkably efficient and accurate, particularly when the number of variables is high.

A. Business understanding

Analysing information that includes text is important for analysing qualitative or quantitative datasets, like knowing the preferences of the people just analysing a chat, a conversation, a review, or a pot, in this case, the purpose is to recognize if the reviews of the videos contain more opinions or complain, so on, at the end could be easy classifier any review.

B. Data understanding

In the TikTok dataset, there is one variable, that contains opinions or claims about the videos, for that reason, these columns are stored in a table for analysis, keeping all the observations and the columns claim_status and video_transcription_text. Keeping, 19,084 observations and 2 columns.

C. Data Preparation

The Data Preparation for this technique was based on the notes in the Naïve Bayes Laboratory class. The steps followed in text mining are:

1. Transformed data by "tm" package in R to create a volatile corpus that contains the "text" vector.
2. Check if the information was stored in the vector.
3. Transforming all lowercase letters in the same format.
4. Removing the numbers, conjunctions, articles, and punctuation.
5. Performing the words to a verb word, it says, if the text contains words like "liking", "likes", like" to "like".
6. Removing white space from the document.

D. Model Building

Split the dataset into training and testing datasets and save their vector label rows. First, to build the model, it created a visualization by package "wordcloud":

- Visualization of the most 50 frequent words[3].

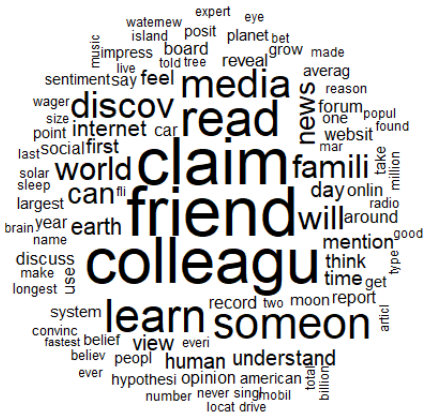


Figure 10. The 50 popular words

Moreover, it pre-processed the data a bit more, like removing words that appear less than 5 times, replacing values greater than zero with yes, and values not greater than zero with. Finally, it was built the model applying the Naïve Bayes Classifier.

E. Test and Evaluation

Using the test dataset that contains 4,771 rows, the model was tested, and the predictions are in 163 claims and 4, 608 opinion reviews.

```
   Cell Contents
|-------------------------|
|                       N |
|-------------------------|


Total Observations in Table:  4771


              | tik_test_labels
tik_test_pred |   opinion | Row Total |
--------------|-----------|-----------|
        claim |       163 |       163 |
--------------|-----------|-----------|
      opinion |      4608 |      4608 |
--------------|-----------|-----------|
 Column Total |      4771 |      4771 |
--------------|-----------|-----------|
```

Figure 11. Predictions by Naïve Bayes Classifier

[3] The code file includes two more visualizations, comparative visualizations between claim and opinion.

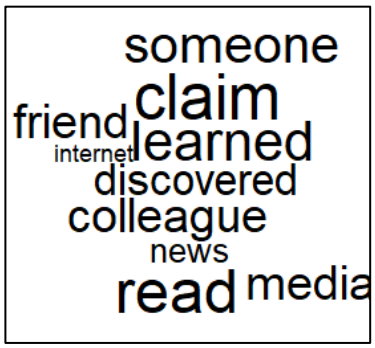


Figure 12. top ten of the words in claim reviews.

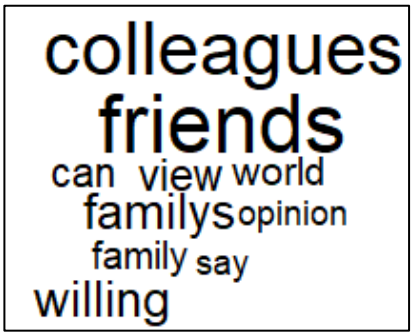


Figure 13. Top ten of the words in opinion reviews

VII. LOAN DATASET

Multiple Linear Regression

Multiple Linear Regression (Pardoe, 2020) is a technique that helps to know if there are relationship between a dependent variable and one or more predictor variables, if there is an association with one of the independent variables, the predictor variable will be influenced.

Gauss-Markov Assumptions

- Correct functional form in the model.
- Homoscedasticity, means, variance is constant.
- No Autocorrelation between errors
- Predictor variables must be independent of the error term.
- It assumes that errors are normally distributed.
- It assumes there is no multicollinearity.
- There are no influential data points.

Evaluating a model (Huy Hoang Nguyen, 2023) means verifying that the model assumptions are met for obtaining the outcomes, these assumptions are the next:

1. A linear relationship between the dependent variable and exploratory variables –by scatter plots of each variable vs the dependent variable or The Durbin-Watson test-
2. Normality of the residuals. –by histogram or QQ Plot-
3. Homoscedasticity of the residuals.-by Scatter plot of Predict Value vs Model Residual-
4. Independent samples.

No or little Multicollinearity.- by Variance Inflation Factor (VIF).

A. Business Understanding

Analyse if the categorical variables and numeric variables are correlated for predicting the loan amount by applying Multiple Linear Regression, and make sure if the accuracy of the model and predictions help to trust in the model.

B. Data Understanding

The Loan file from The Kaggle Website contains 38,576 observations and 24 columns, which has the variables id, address_state, application_type, emp_lenght, emp_tittle, grade, home_ownership, issue_date, last_credit_pull_date, last_payment_date, loan_status, next_payment_date, member_id, purpose, sub_grade, term, verification_status, annual_income, dti, installment, int_rate, loan_amount, total_acc, total_payment.

Four of the variables are datetime, eleven categorical variables and seven are numeric variables. It will display some visualizations of the information in the dataset.

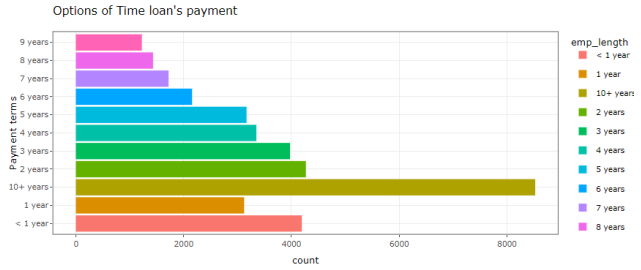


Figure 14. Bar plot of payment terms

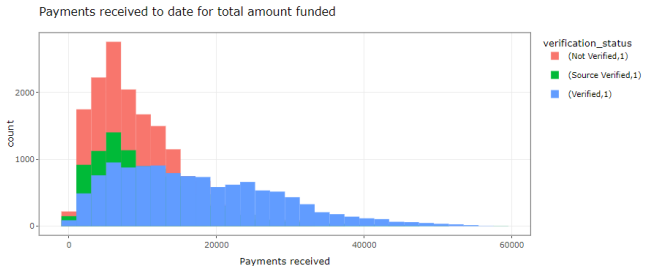


Figure 15. Histogram of payment received.

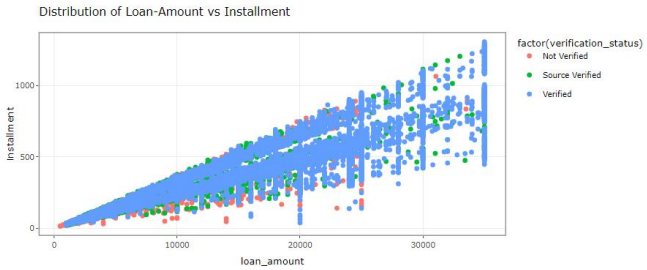


Figure 16. Scatterplot of loan_amount vs installment.

C. Data Preparation

- Move the position of the column "id" in the RowNames.

- Checking unique values of the columns, it find that the variable "application_type" has one only value called "INDIVIDUAL", so, it has been removed.
- Convert the variables containing a date to "DateTime" type.
- It found 55, 233 missing values in the dataset, it says, 7% of the values. These values are deleted from the dataset.

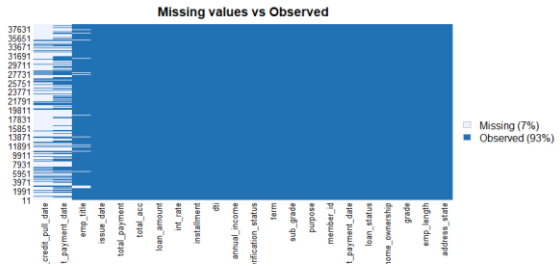


Figure 17. Missmap of missing values

- Handling outliers by visualizations and interquartiles, it found outliers in the variables annual_income, dti, installment, int_rate, loan_amount, total_acc and total_payment, all these outliers had been removed from the dataset.

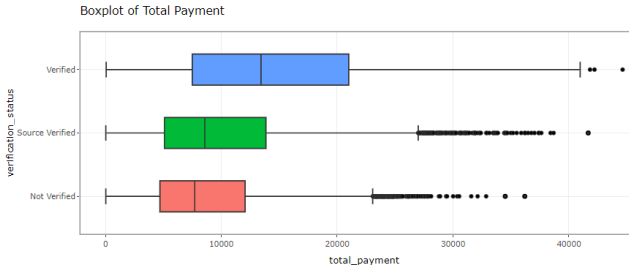


Figure 18. Boxplot of total_payment

- Having a view of the numeric variables, it checked by correlogram the relationship between the variables and it found just a relationship between loan_amount, installment, and total_payment.

Variable	Mean	Median	STD Deviation	Skewness	Kurtosis
annual_income	5.87E+04	54500	2.60E+04	0.706082	3.045257
dti	1.34E-01	0.135	6.67E-02	-0.04298929	2.147822
installment	2.81E+02	257.24	1.57E+02	0.71298658	3.01749
int_rate	1.17E-01	0.1166	3.53E-02	0.22506646	2.440237
loan_amount	9.59E+03	8600	5.53E+03	0.66666242	2.778413
total_acc	2.05E+01	19	9.84E+00	0.49383983	2.599783
total_payment	1.02E+04	8960	6.39E+03	0.68732501	2.724459

Table 1. Descriptive statistics

- By the visualizations and and descriptive statistic in the table 1, the variable annual_income and installment were transformed by LogNatural[4].
- Performing the categorical variables by dummyVars function and Hot-Encoding the variables: home_ownership.NONE, home_ownership.OTHER, home_ownership.OWN, home_ownership.RENT, loan_status.Current, loan_status.Fully.Paid, term..60.months, verification_status.Source.verified and verification_status.Verified.

D. Model Building

The dataset is split into train and test datasets, and the predictor variable is loan_amount,. There were built 4 Models, the first one was applying the equation in figura 19. The second Model removed la variable emp_lengt because the P-value >0.05, in the third model, it was removed the variables dti[5].

```
lm(formula = loan_amount ~ emp_length + annual_income + dti +
    installment + int_rate + total_acc + total_payment + home_ownership.NONE +
    home_ownership.OTHER, data = train, subset = -home_ownership.OWN +
    home_ownership.RENT + loan_status.Current + loan_status.Fully.Paid +
    term..60.months + verification_status.Source.Verified + verification_status.Verified)
```

Figure 19. First Model's equation.

The response variable loan_amount is explained by the variables installment and total_payment in the fourth Model, any of the categorical variables participates in the value of the loan_amount. The outcome of the model in R is showing an accuracy of 86.78%. and each of the variable has a P-value under 0.05.

```
Call:
lm(formula = loan_amount ~ installment + total_payment, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-4713.7  -928.7  -502.8   177.1 20016.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.810e+02  2.766e+01   13.78   <2e-16 ***
installment   2.307e+01  1.576e-01  146.39   <2e-16 ***
total_payment 2.659e-01  3.891e-03   68.33   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2012 on 22260 degrees of freedom
Multiple R-squared:  0.8678,    Adjusted R-squared:  0.8678
F-statistic: 7.305e+04 on 2 and 22260 DF,  p-value: < 2.2e-16
```

Figure 20. Fourth Model of Multiple Linear Regression

E. Test and Evalution

The Normal probability plot (Figure in the top right) shows that the relationship between the sample percentiles and the theoretical percentiles is not linear, the condition that the error terms are normally distributed is not met.

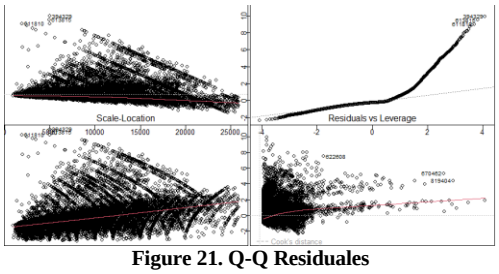


Figure 21. Q-Q Residuales

[4] See the visualizations of the Normal distribution plot and Q-Q plot in the Loan Code file.

[5] Each model is displayed step by step in the Loan code File

K-Nearest Neighbors Classifier

The task of the K-Nearest Neighbors Classifier (Anthony So, The Applied Artificial Intelligence Workshop, 2020) is to determine which class data points belong to which group, by dividing data like the goal of classification algorithms. On the other hand, the neighbors are chosen from a set of objects for which the class or object property value is known, this is applied in both cases, K-NN classification or regression.

The proximity measure that calculates the distance between data points is the base of the K-nearest neighbors algorithm to find the relevant relationship or patterns between the features that will lead to this class. Euclidean distance (Sabry, 2022) is a typical distance metric for continuous variables, another metric, that can be used for discrete variables is Hamming distance.

Often learning the distance metric with specific algorithms like Large Margin Nearest Neighbor or Neighborhood Components Analysis can increase the classification of accuracy of k-NN.

A. Business Understanding

Requesting a financial loan implies covering a variety of requirements, and a long time until it is approved, in the case, it will be approved. Therefore, knowing what those requirements are, will make it easy to cover them. The purpose of applying the K-NN Algorithm is to know the accuracy of asking the status of verification of the loan, which means if it has been verified, not verified or just source verified. Furthermore, analyze what requirements have a big weight in the approval of selecting the term of payments for 36 months or 60 months.

B. Data Understanding

The variables will be analysed in K-NN algorithm are verification_status, term, annual_income, dti, installment, int_rate, loan_amount and total_payment, the label or target variables, in the first model, the dependent variable is verification_status and in the second model is replace for the variable term.

C. Data Preparation

In view of the fact that the algorithm calculates the Euclidean distance from k value to data points and it is sensitive to magnitudes, all the features were scaled to weight in equally.

The values of the targets that will be predicted, were re-write again without spaces or numbers in the beginning of the names (i.e., 30 months to months_30 or Not Verified to Not_Verified).

The dataset was split into training and testing by the *createDataPartition* function.

D. Model Building

Considering that the algorithm works by making iterations from end to end of the searching the k-value from the range of k-value and checking model accuracy in every search, it was created the *control* function for 10 iterations and repeated the cross-validation 3 times. Defining the target by the hypothesis that the status loan has been verified, not verified or source verified, built the model 1 and following the Kappa accuracy, which is low than 0.20, the model is rejected as none to slight.

```
k-Nearest Neighbors

22227 samples
    7 predictor
    3 classes: 'Not_Verified', 'Source_Verified', 'Verified'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 20003, 20004, 20005, 20003, 20005, 20005, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.4617062  0.1323525
  7  0.4763434  0.1475125
  9  0.4871407  0.1575075

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 9.
```

Figure 22. K-Nearest Neighbors Model 1

The second model by the hypothesis that the months of the loan's payment depend on the independent variables, i.e., to pay for 36 or 60 months, in each iteration of the algorithm, the ROC's accuracy is explained over 95%, it says, the relation with the True Positive Rate and False Positive Rate is good with 9 neighbors.

```
k-Nearest Neighbors

22226 samples
    7 predictor
    2 classes: 'months_36', 'months_60'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 20002, 20003, 20003, 20003, 20004, 20004, ..
Resampling results across tuning parameters:

  k  ROC        Sens       Spec
  5  0.9562912  0.9885711  0.7650062
  7  0.9621460  0.9896923  0.7476476
  9  0.9642286  0.9910693  0.7312992

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 9.
```

Figure 23. K-Nearest Neighbors Model 2

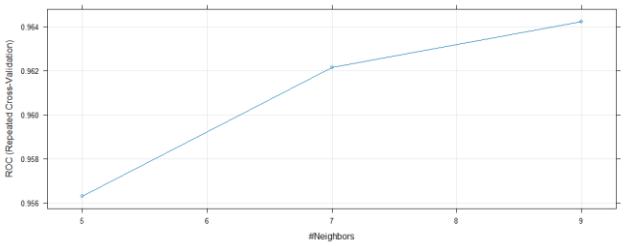


Figure 24. Plot of Elbow for training dataset

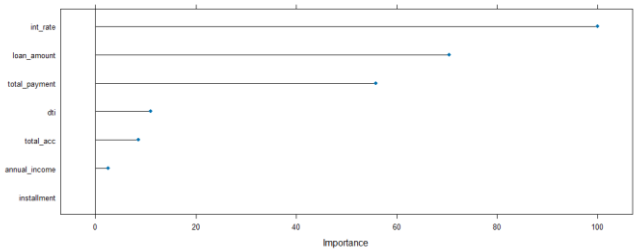


Figure 25. Importance of each variable in the Model 2

The figure above (25) shows that int_rate, loan_amount, and total_payment will influence whether the number of monthly payments is for 36 or 60 months.

E. Test and Evaluation

The prediction of the model presents 93.09% accuracy, both true positives and true negatives among the total number

of cases, the Kappa metric indicates a good agreement between the predictions and true values (0.79).

The sensitivity is 99.10%, it says, the proportion of 36 months cases was correctly identified by the model and the proportion of 60 months cases was negative by the model with a specificity of 73.78%.

```
Confusion Matrix and Statistics

          Reference
Prediction  months_36 months_60
  months_36      7197       593
  months_60        65      1669

               Accuracy : 0.9309
                 95% CI : (0.9256, 0.9359)
    No Information Rate : 0.7625
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7926

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9910
            Specificity : 0.7378
         Pos Pred Value : 0.9239
         Neg Pred Value : 0.9625
             Prevalence : 0.7625
         Detection Rate : 0.7557
   Detection Prevalence : 0.8179
      Balanced Accuracy : 0.8644

       'Positive' Class : months_36
```

Figure 26. Confusion Matrix of K-NN Model1

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  9524


                    | prediction
dfk2[-splitt3, 1] | months_36 | months_60 | Row Total |
------------------|-----------|-----------|-----------|
        months_36 |      7197 |        65 |      7262 |
                  |     0.991 |     0.009 |     0.762 |
                  |     0.924 |     0.037 |           |
                  |     0.756 |     0.007 |           |
------------------|-----------|-----------|-----------|
        months_60 |       593 |      1669 |      2262 |
                  |     0.262 |     0.738 |     0.238 |
                  |     0.076 |     0.963 |           |
                  |     0.062 |     0.175 |           |
------------------|-----------|-----------|-----------|
     Column Total |      7790 |      1734 |      9524 |
                  |     0.818 |     0.182 |           |
------------------|-----------|-----------|-----------|
```

Figure 27. Cross-Table of Prediction

VIII. AirBnB Dataset

Decision Tree

Tree-based models (Anthony So, The Appplied Artificial Intelligence Worshop, 2020) are very popular as they can model complex non-linear patterns and are easy to interpret. A decision tree consists of rules to formulate a decision on the prediction of a data point. Every edge of an internal node represents a possible value or a possible interval of values of the tree while every node of the decision tree represents a feature.

One of the frequent problems for making a decision based on a few data points is overfitting in the decision tree. It makes sense to define an efficient way of constructing a decision tree since it can build it in any possible order. Consequently, entropy measures are the highest possibility of an event, the higher the degree of randomness is the higher the entropy of the attribute.

A. Business Understanding

Airbnb is like an online marketplace for short and long-term homestays and experiences, the purpose of applying the decision tree in this section is to predict what type of room will be booked according to the price, minimum of nights, number of reviews, reviews per month, calculated host listing counts, availability 365 and number of reviews ltm.

B. Data Understanding

The third dataset is about Airbnb's listings, which stores 14,861 rows and 18 columns, four of them are categorical data, two are logical, and four are twelve are numeric data. The variables name are: id, name, host_id, hots_name, neighbourhood_group, neighbourhood, latitude, longitude, room_type, price, minimum_nights, number_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365, number_of_reviews_ltm and license.

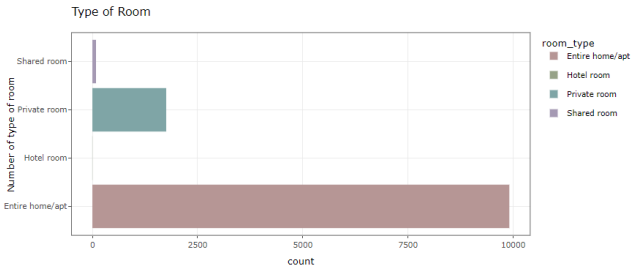


Figure 28. Bar plot of type of room

Figure 29. Bar plot of type of room

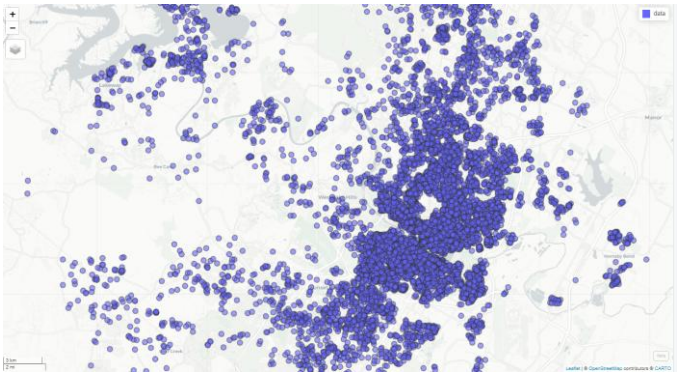


Figure 30. Distribution of the AirBnB in San Antonio, Texas

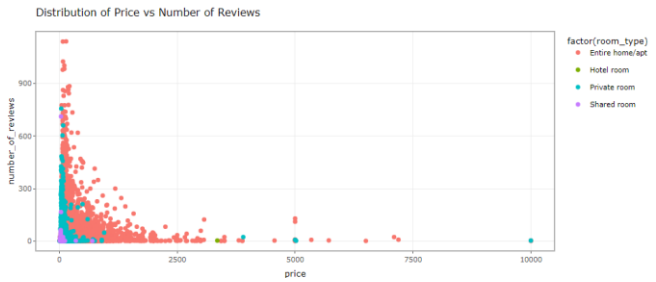


Figure 31. Dispersion plot of Price vs number of reviews

C. Data Preparation

The "id" variable is moved to column "RowNames" as the index variable; checking the unique values, it saw that the variable neighborhood and license had the same value in each row, for that reason, they were deleted. Other important preprocessing processes are:

- Handling missing values, only 3% of the missing values were found and removed.

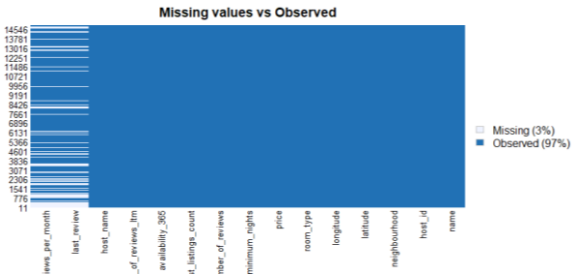


Figure 32. Missmap of missing values

- There was existence of outliers in the variable Price, the minimum_night is a special case because values from O to 27 minimum of nights and atypical values over 300 minimum nights. Reviews_per_month had numerous outliers, reviews_per_month, calculated_host_listings_count and number_of_reviews_ltm. All of the outliers were deleted.

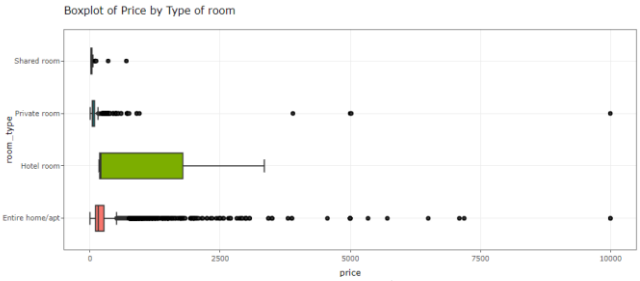


Figure 33. Boxplot of Price variable

- The correlogram shows that there is a correlation between 2 variables, reviews_per_month and number_of_reviews_ltm, the rest of them have a poor correlation between them.
- Having a review of the variables, it defines as target the variable room_type and nodes of the tree decision are defined for the independent variable: price, minimum_nights, number_of_reviews, reviews_per_month, calculated_host_listings_count, availability_365 and number_of_reviews_ltm.
- Splitting dataset into training and testing by the *createDataPartition* function.

D. Model Building

It designed the model like response variable room_type with an outcome in the first node that there is a probability of 100%, the customer decides to book an entire home or apartment, if the price is under 77, 81% probability a customer books a listing, however, if the price is over 76, there is a probability of 19%, the listing will be booked.

Following that the price is over 76, there is a probability that 9% of customers, will book a private room if the price is less than 58, by this condition there is a probability of 2% the listing will be booked for less than 9 nights, and it is an entire home or apartment. On the other hand, there is a probability that a customer who books for more than 8 nights, it will book a private room at 7% of probability.

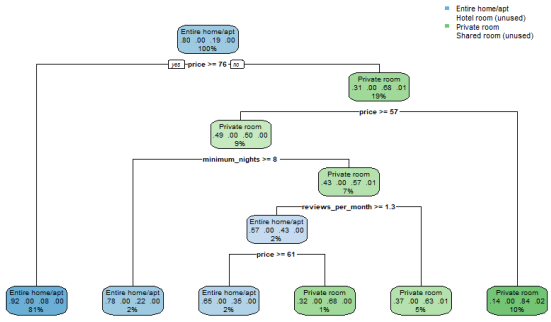


Figure 34. Decision tree Model

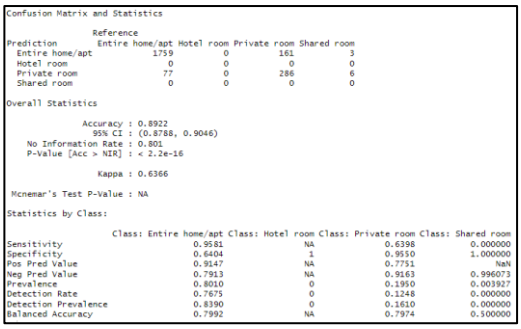
```
Classification tree:
rpart(formula = room_type ~ ., data = training, method = "class",
    xval = 20)

Variables actually used in tree construction:
[1] minimum_nights    price             reviews_per_month

Root node error: 1067/5352 = 0.19936

n= 5352
```

Figure 35. Decision tree Model

Figure 34 shows the variables used in the building of the model.

E. Test and Evaluation

The accuracy of the predictions is 89.22%, providing a more balanced assessment of the model's performance, Kappa metric indicates an agreement between the predictions and the true values with a value of 0.6366. The confusion Matrix displays statistics by class, i.e., the entire home/ apartment has a sensitivity value of 0.9581 versus the private room, which has a sensitivity value of 0.6398, explaining that there is more proportion to book an entire home/ apartment than a private room, on the other hand, to book a hotel room does not have any probability that will be booked and in the case to book a shared room, the probability is not zero but neither significant.



```
Confusion Matrix and Statistics

                Reference
Prediction      Entire home/apt Hotel room Private room Shared room
  Entire home/apt            1759          0          161          3
  Hotel room                    0          0            0          0
  Private room                 77          0          286          6
  Shared room                   0          0            0          0

Overall Statistics

               Accuracy : 0.8922
                 95% CI : (0.8788, 0.9046)
    No Information Rate : 0.801
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6366

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Entire home/apt Class: Hotel room Class: Private room Class: Shared room
Sensitivity                          0.9581               NA              0.6398           0.000000
Specificity                          0.6404                1              0.9550           1.000000
Pos Pred Value                       0.9147               NA              0.7751                NaN
Neg Pred Value                       0.7913               NA              0.9163           0.996073
Prevalence                           0.8010                0              0.1950           0.003927
Detection Rate                       0.7675                0              0.1248           0.000000
Detection Prevalence                 0.8390                0              0.1610           0.000000
Balanced Accuracy                    0.7992               NA              0.7974           0.500000
```

Figure 36. Confusion Matrix of Decision Tree

IX. Conclusion and Future Wok

There is a vast knowledge of the application of data mining and machine learning, which help to solve different areas like finance, social media, marketing, hospitality, etc. However, is important to be clear that analyzing information does not always follow the same pattern, the presence of issues depends on each company and the requirements that they are looking for. Since each dataset contains different information, that does not mean that every algorithm is applicable, or the accuracy will be good. The way to get more experience in the understanding of these areas, it practices and study because the theory is not enough, and practice without theory as well is not enough, having each involved concept in the analysis is important, because, it means, applying data mining and modelling, makes you feel confident about what you are doing and do no do more and more job, taking to you more time. Additionally, knowing a programming language is completely important, it was the main issue doing this project, thankfully the laboratory classes were helpful.

References

[1] Anthony So, W. S. (2020). The Applied Artificial Intelligence Workshop. Packt Publising.

[2] Anthony So, W. S. (2020). The Appplied Artificial Intelligence Worshop. Packt Publishing.

[3] David W. Hosmer Jr, S. L. (2013). Applied Logistic Regression 3rd Edition. Wiley.

[4] Delen, D. (2014). Real-World Data Mining: Applied Business Analytics and Decision Making. Pearson.

[5] Gary D. Miner, J. E. (2012). Practical Text Mining and Statistical Analysis for Nonstructured Text Data Applications. Academic Press.

[6] Huy Hoang Nguyen, P. N. (2023). "Building Statisticasl Models in Python". UK: Packt Publishing.

[7] Hyken, S. (2022). Your Business' Guide to TikTok Marketing. Forbes.

[8] Michelli, J. (2019). The Airbnb Way: 5 Leadership Lessons for Ignithing Growth Through Loyalty, Community, and Belonging. McGraw-Hill.

[9] Pardoe, I. (2020). "Applied Regression Modeling". Wiley.

[10] Reinhard, M. (2010). Commercial Mortgages 101: Everything you need to know to Create a Winning Loan Request Package. AMACOM.

[11] Rohit Raja, K. K. (2022). Data Mining and Machine Learning Applications. Wiley-Scriver.

[12] Sabry, F. (2022). K Nearest Neighbor Algorithm. One Billion Knowledgeable.

[13] https://www.kaggle.com/datasets/yakhyojon/tiktok

[14] https://www.kaggle.com/datasets/deeplearner09/airbnb-listings

[15] https://www.kaggle.com/datasets/nezukokamaado/auto-loan-dataset