Machine Learning Algorithms in R for House Price Prediction and Credit Default Classification

Alfredo
University of London
ST3189 - Machine Learning
Dr Chew Jee Loong
April 3, 2025

University of London Student's ID: 240724685
Programme: Graduate Diploma in Business Analytics
University of London Student's E-mail: ma1679@london.ac.uk
Learning Institution: Singapore Institute of Management

Word count: 2989 (Exclude title page and Table of content)
Number of pages: 10 (Exclude title page and Table of content)

Table of Contents

King’s County House Price Prediction

1.1 Background

King’s County, Seattle, Washington, is an excellent location for investment. Based on an article in King and Snohomish Counties (2025), King Counties have grown significantly in population, with Forbes ranked this area to be the second in fastest growing city in the US.

This section will focus on prediction for housing prices and evaluate whether these predictions remain valid in current market. While the real world may have complex interactions, Kaushik (n.d) mention that linear regression is a reliable start to capture and understand the primary relationships in house price dataset.

The dataset used in this analysis will be from an open-access domain, an OpenML website (OpenML, n.d), https://www.openml.org/search?type=data&status=active&id=44989 This dataset contains house sale prices for King County between May 2014 and May 2015.

1.2 Research Question

This report will be conducted to answer these research questions:

1. What variables affect housing prices in King’s County?
2. How good is the model in predicting current prices utilizing historical data?

1.3 Exploratory Data Analysis

1.3.1 Data Preparation

The dataset from King’s County have 21614 observed data that includes 15 input variables, which are: number of bedrooms, number of bathrooms, square feet of living room, square feet of lot, floor levels, waterfront grade, view grade, condition, house grade, square feet of basement, year built, year renovated, zip code, latitude and longitude, with its predicted variable house price.

This process shows that there is one missing value or duplicated data, subsequently reducing the dataset into 21613 number of rows. Etherington, T. R. (2021) reported, The Mahalanobis distance is a statistical technique that has been used in statistics and data science for data classification and outlier detection, Mahalanobis functions effectively on multivariate data considering it utilizes covariance matrix of variables to find the length between data points and the middle point. Total rows are reduced up to 19,900 by implementing this function to identify and remove possible outliers. This step is executed by realizing the effect of noise and outliers could have for the model fit in future analysis.

1.3.2 Correlation Matrix

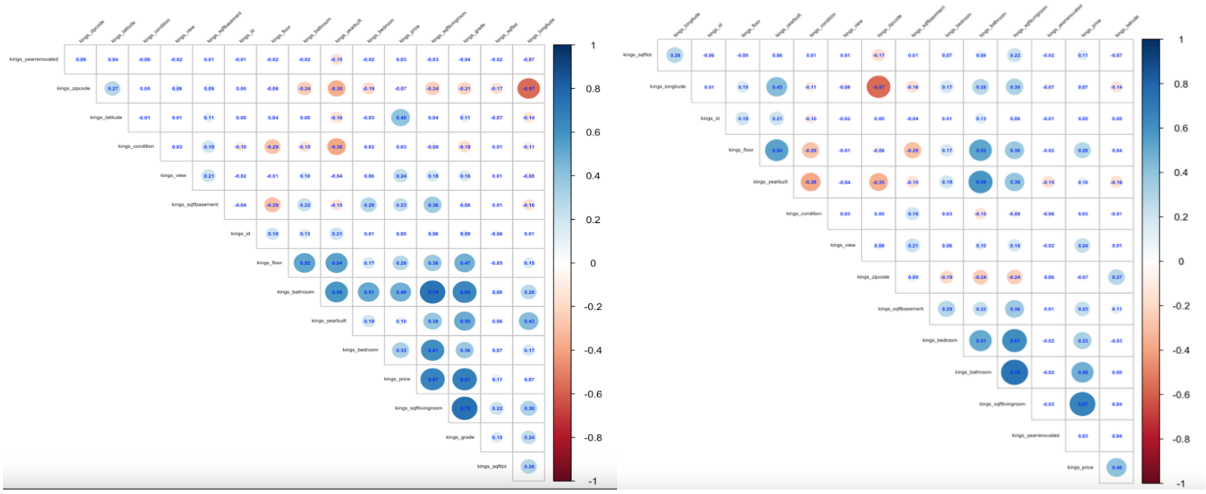


Figure 1.3.2 Correlation matrix before and after exclusion of house grading

This correlation matrix shows housing price has positive correlation of 0.67 with square feet of living room, it is the same as housing grade. Number of bathrooms also positively correlates with square feet of living room, which is ideal as the size of the room increases, there is a high potential of having more bathrooms. Multicollinearity primarily concerns around the grade feature, keeping the grade feature will possibly mislead the model into relying into subjective input such as grading rather than on factual data such as square feet of living room.

According to Tilii7 (2020), Multi-collinearity features could skew the variance and only contain useless information. Deleting one of the highly correlated feature with certain threshold is the best way of dealing with multi-collinearity. Zhang et al. (2020) added, minimal research has been done to confirm the effectiveness of user-defined labels, therefore, removing grade will be appropriate as subjective grading might contain bias from data entry.

1.3.3 Scatterplot & trendline

In reference to Nau (n.d), there are four important assumptions to justify linear regression model usage; Relationship between feature and response should be linear, independent residual, homoscedasticity and the error should be normally distributed. These assumptions confirm the use of linear regression so that the model can predict accurately, unbiased and statistically valid and finally leading to the correct inferences.

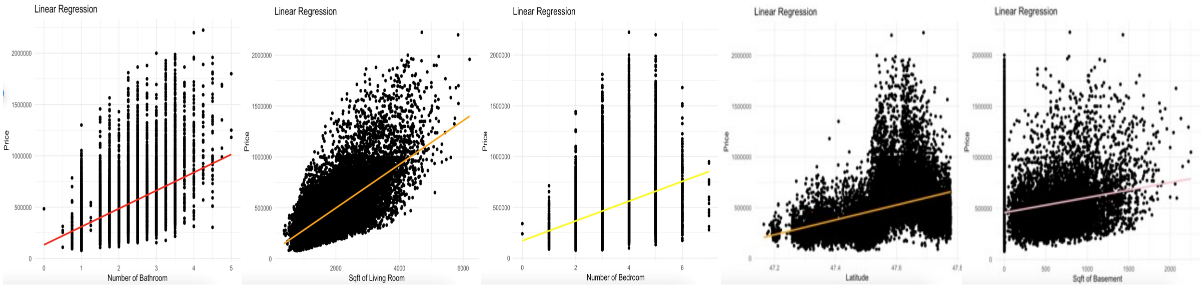


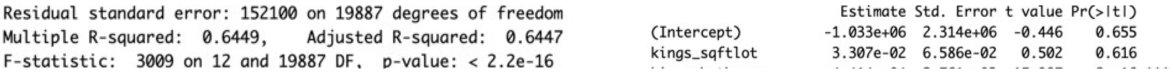
Figure 1.3.3 Scatter plotting observations to visualize data pattern

By observing these scatterplots, square feet of living room and number of bedrooms has an upward trend and linear pattern against housing prices, one of the mild concerns would be the rule of constant variance as there is a significant spread in the middle area for number of bathroom and increasing spread for square feet of living room. Reported by Penn State University (n.d), if homoscedasticity is violated, a large number of samples could justify mild violation. Alternatively, weighted least squares can be used to address such problem. It is important to verify whether the variance is indeed non-constant or if the observed pattern is due to overlapping data points, which may obscure the visualization of the entire dataset.

Number of bedrooms showing a convex data pattern trend that has a peak in 4 bedrooms that is starting to decline afterward, there is also a slight increase of housing prices starting from 4 bedrooms and onward. Convex data pattern also applies to latitude, there is a sudden peak at latitude around 47.6 point. Application of non-linear regression will be appropriate. Polynomial regression is one statistical method suitable for curved pattern on, as confirmed by MathMonks (2024).

The data distribution for square feet of basement shows a general agreement on upward trend, one major problem is the absence of basement feature combines directly houses without basement with 0 input and houses with basement with its respective basement size. According to ProjectPro (2024), splitting features into different part will likely improve the value of the features toward the response. Since the data points in no basement are concentrated in 0 axis, employing standardizing technique directly will not capture the true relationship of this feature as intended. Hence, feature splitting will be conducted.

1.4 Baseline Model



```
Residual standard error: 152100 on 19887 degrees of freedom
Multiple R-squared:  0.6449,	Adjusted R-squared:  0.6447
F-statistic:  3009 on 12 and 19887 DF,  p-value: < 2.2e-16
```

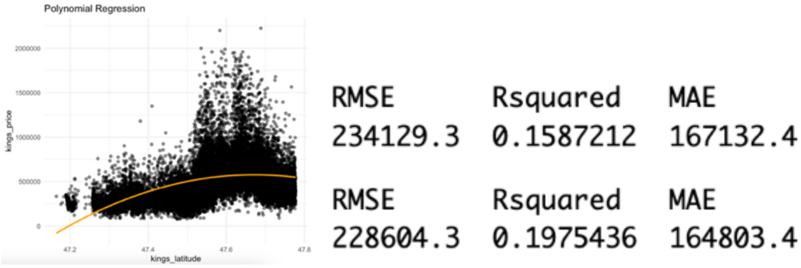
```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.033e+06  2.314e+06  -0.446    0.655
kings_sqftlot    3.307e-02  6.586e-02   0.502    0.616
```

Figure 1.4.1 Baseline model for comparison

Fitting simple linear model using all the observations and features, most of the features are statistically significant shown with a low P-value, also notice the value of R^2 explain the model’s ability to capture approximately 64.5% of variances in the dataset.

1.5 Feature Engineering

1.5.1 Polynomial Transformation

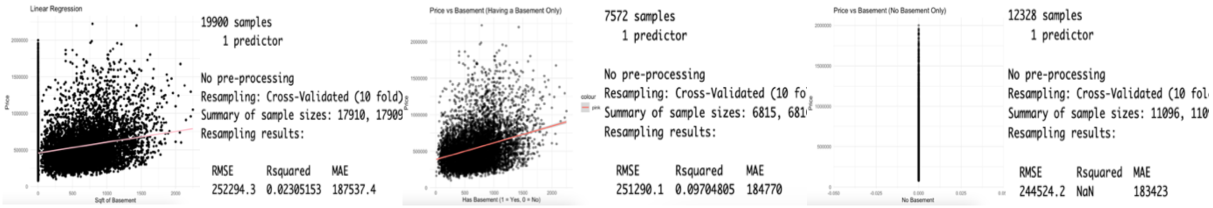


```
RMSE      Rsquared   MAE
234129.3  0.1587212  167132.4

RMSE      Rsquared   MAE
228604.3  0.1975436  164803.4
```

Figure 1.5.1 Polynomial transformation for latitude

Implementing polynomial regression 2nd degree for latitude increase R^2 around 4% while also decreasing RMSE. Even it fits better, one major problem is trendline map out a negative turn for latitudes greater than 48.2. This will constraint model's ability to generalize outside latitude roughly <47.2 and 48.4>. Another report from Penn State University (n.d), extrapolating beyond the scope of model may be risky. In summary, latitude feature has undergone 2nd degree transformation which only applicable for analysis that is in the scope of this study.

1.5.2 Feature Splitting



```
19900 samples
    1 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 17910, 17909
Resampling results:

  RMSE      Rsquared    MAE
  252294.3  0.02305153  187537.4
```

```
7572 samples
   1 predictor

No pre-processing
Resampling: Cross-Validated (10 fo
Summary of sample sizes: 6815, 681
Resampling results:

  RMSE      Rsquared    MAE
  251290.1  0.09704805  184770
```

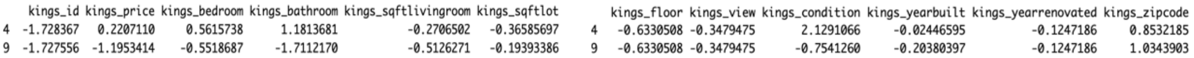
```
12328 samples
    1 predictor

No pre-processing
Resampling: Cross-Validated (10 fol
Summary of sample sizes: 11096, 110
Resampling results:

  RMSE      Rsquared  MAE
  244524.2  NaN       183423
```

Figure 1.5.2 Feature splitting for basement

Adopting feature splitting to construct two other independent features; with basement and no basement, there is a significant increase of R^2 into 9.7% for with basement but subsequently a NaN R^2 for no basement. It is an unsurprising value as no basement provide same data for every row which is 0. Since R^2 = 0, the estimated regression line will be an ideal horizontal. The feature explained no variation of our response, (Penn State University, n.d).

9.7% is considered medium to high amount of information explained. This shows an importance of using splitted square feet of basement feature to elevate the accuracy of prediction. Although the sample size now fall into 7,572, University of Alabama at Birmingham (2021) reported that in order to construct a classification neural network that could ensure at least a specified fraction of images a classified correctly, approximately 4,000 observations per class were needed. Which in this case, sample size of 7,572 will still be sufficient to build a regression model.

1.5.3 Feature Scaling



```
  kings_id kings_price kings_bedroom kings_bathroom kings_sqftlivingroom kings_sqftlot
4 -1.728367   0.2207110     0.5615738      1.1813681           -0.2706502  -0.36585697
9 -1.727556  -1.1953414    -0.5518687     -1.7112170           -0.5126271  -0.19393386
```

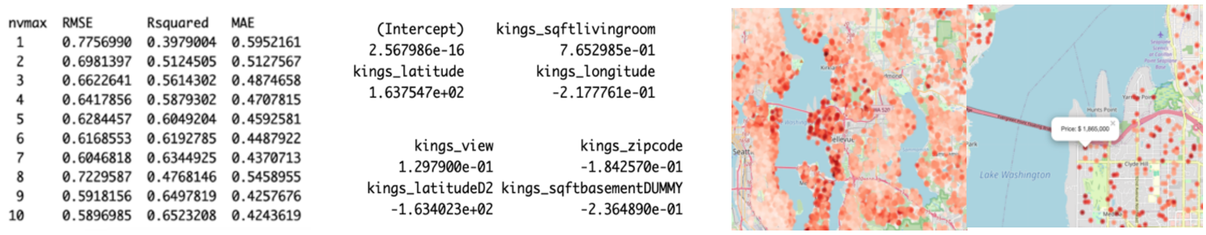
```
  kings_floor kings_view kings_condition kings_yearbuilt kings_yearrenovated kings_zipcode
4  -0.6330508 -0.3479475       2.1291066     -0.02446595          -0.1247186     0.8532185
9  -0.6330508 -0.3479475      -0.7541260     -0.20380397          -0.1247186     1.0343903
```

Figure 1.5.3 Consistent feature importance with feature scaling

Observing the range of value for housing prices which is from 75,000 to 770,000 and some feature that has a small value such as latitude, indicating major problem for biased learning. Utilizing standardization to ensure all variable hold the same weight is crucial because these features deviate extremely unequal from the mean. There is also no perfect method to confirm the accurate feature weightage for current dataset. As stated by Zhang (2024), utilizing standardization will force each coefficient to represent the change in response for one standard deviation away.

1.6 Modelling

1.6.1 Subset selection



```
nvmax  RMSE       Rsquared   MAE
 1     0.7756990  0.3979004  0.5952161
 2     0.6981397  0.5124505  0.5127567
 3     0.6622641  0.5614302  0.4874658
 4     0.6417856  0.5879302  0.4707815
 5     0.6284457  0.6049204  0.4592581
 6     0.6168553  0.6192785  0.4487922
 7     0.6046818  0.6344925  0.4370713
 8     0.7229587  0.4768146  0.5458955
 9     0.5918156  0.6497819  0.4257676
10     0.5896985  0.6523208  0.4243619
```

```
    (Intercept)    kings_sqftlivingroom
   2.567986e-16            7.652985e-01
kings_latitude          kings_longitude
   1.637547e+02           -2.177761e-01


         kings_view           kings_zipcode
       1.297900e-01           -1.842570e-01
  kings_latitudeD2 kings_sqftbasementDUMMY
      -1.634023e+02           -2.364890e-01
```

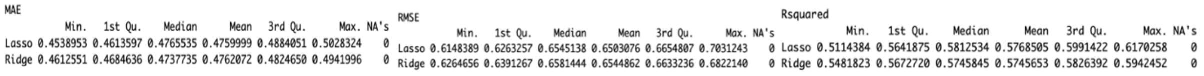
Figure 1.6.1 Selecting key features to build model

According to Hastie, Tibshirani, & Tibshirani (2017), signal-to-noise ratio could serve as an indicator when choosing subset selection or regularization. Best subset selection performs best in high SNR regimes; in contrast, Lasso did better in low SNR regimes. Since dataset's signal-to-noise ratio is unknown, applying lower-risk approach by implementing subset selection first is expected.

Maximum number of predictors 7 shows 63.44% R^2 in which at 8 predictors there is a significant drop to 47.68% R^2 and quickly rise again at 9 predictors to 64.97% R^2. Model will stop adding predictors after 7 to balance smaller number of predictors yet a good R^2 %. The middle figure showing additional insights: square feet of living room reveal a large explanatory power R^2 39.79%, follow up by view grade ~11% and the other predictor are mostly geographical except for square feet of basement. Inferring significant contribution of house sizes and house location, displayed in heatmap figure.

Last figure on the right-side display how geographical location of the house affects housing price, amplifying the value of those features. Showing the dependencies of every geographical feature, together increasing or decreasing house prices.

1.6.2 Regularization



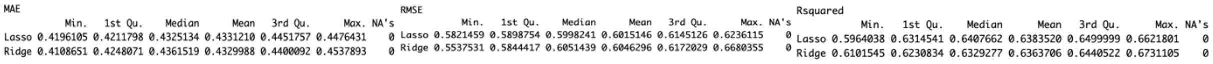
```
MAE
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Lasso 0.4538953 0.4613597 0.4765535 0.4759999 0.4884051 0.5028324    0
Ridge 0.4612551 0.4684636 0.4737735 0.4762072 0.4824650 0.4941996    0
```

```
RMSE
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Lasso 0.6148389 0.6263257 0.6545138 0.6503076 0.6654807 0.7031243    0
Ridge 0.6264656 0.6391267 0.6581444 0.6544862 0.6633236 0.6822140    0
```

```
Rsquared
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Lasso 0.5114384 0.5641875 0.5812534 0.5768505 0.5991422 0.6170258    0
Ridge 0.5481823 0.5672720 0.5745845 0.5745653 0.5826392 0.5942452    0
```

Figure 1.6.2.1 Underfitting application of regularization after subset selection

Since some high coefficient such as 4.06 view grade after standardization still exists, the model will be penalized further to achieve better fit. These figure displays that after penalization the model underfitted. Thus, regularization for subset selection won’t be applied.



```
MAE
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Lasso 0.4196105 0.4211798 0.4325134 0.4331210 0.4451757 0.4476431    0
Ridge 0.4108651 0.4248071 0.4361519 0.4329988 0.4400092 0.4537893    0
```

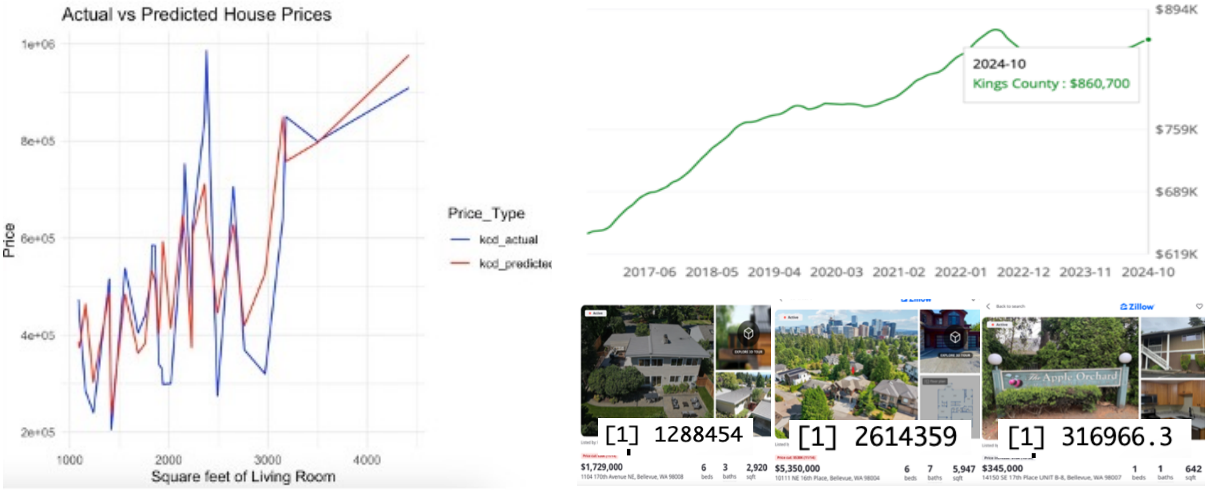
```
RMSE
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Lasso 0.5821459 0.5898754 0.5998241 0.6015146 0.6145126 0.6236115    0
Ridge 0.5537531 0.5844417 0.6051439 0.6046296 0.6172029 0.6680355    0
```

```
Rsquared
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Lasso 0.5964038 0.6314541 0.6407662 0.6383520 0.6499999 0.6621801    0
Ridge 0.6101545 0.6230834 0.6329277 0.6363706 0.6440522 0.6731105    0
```

Figure 1.6.2.2 Comparing best subset selection and regularization

Using best subset selection at 7 subsets R^2 = 63.44% and RMSE = 60.46%, while Lasso achieves better result at R^2 = 63.83% and RMSE = 60.15% indicating a minor difference. Since this is the case, best subset selection will be a preferable risk-averse decision to safeguard against noise.

1.7 Model Evaluation & Prediction



```
[1] 1288454
```

$1,729,000 6 beds 3 baths 2,920 sqft
1104 170th Avenue NE, Bellevue, WA 98008

```
[1] 2614359
```

$5,350,000 6 beds 7 baths 5,947 sqft
10111 NE 16th Place, Bellevue, WA 98004

```
[1] 316966.3
```

$345,000 1 beds 1 baths 642 sqft
14150 SE 17th Place UNIT B-8, Bellevue, WA 98007

Figure 1.7 Utilizing model for predicting current house

Figure on the left shows the predictive power of model by utilizing square feet of living room since it is the key feature, with R^2 of ~63.44%. The blue line is created by leveraging a random data from observations including random scaling for variable. The red line is the prediction of model. In this visualization, the model still able to generalize with the trend throughout 25 partially new constructed data. It is worth noting that the model has a standardized RMSE of 60.46%, inferring almost one standard deviation away from the mean of error is expected.

Model will be tested to predict current housing prices in King’s County. Finding the scale for current median prices in King’s County with the dataset’s median price will provide a parameter to serve as a reference for model’s prediction. It is best to utilize median, as Block Siege (2017) mention in their property research that median is often used as an indicator because it accurately reflects the sample size being used, trend of market, sentiment of customer and the overall market condition.

The first house price is $1,729,000 with 6 bedrooms, 3 bedrooms, 2,920 square feet of living room, latitude 47.6161, longitude -122.1223, zip code 98008. By making use of these data excluding house price, the model predicted the house price as $1,288,454. Model's prediction also unable to keep up and become less accurate when the actual house price increases suggesting large houses experience greater price increase over time.

1.8 Summary

Model is built by leveraging King's County dataset from 2014-2015. The most important feature is square feet of living room with 39.79% R^2 clearly indicated by the 67% positive correlation with housing prices then follow up by the geographical variables. Begin building benchmark model and gain 64.49% R^2, realizing that by using benchmark model will fit all the noise and outliers.

The model achieves ~63.44% R^2 with subset selection, reporting findings of a larger prediction error estimates for large houses. Overall, the model has lower explanatory power ~1% but the dataset dimensionality of goes down from 15 to only 7 input variables.

Credit Default Risk Prediction

2.1 Background

It is crucial to understand which prospect customer might be the next defaulting subject. Akinjole, Shobayo, Popoola, Okoyeigbo, & Ogunleye, (2024) report that an accurate prediction of a borrower defaulting on loans will reduce financial losses and help maintain stability and profitability. A model will be built to classify a customer that tend to default at later stages

The dataset used in this analysis will be from an open-access domain, an OpenML website (OpenML, n.d), https://www.openml.org/search?type=data&status=active&id=43454

2.2 Research Question

This report will be conducted to answer these research questions:

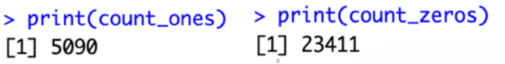
1. What variables contributes to the defaulting rate of customer?
2. How accurate is the model in classifying default status?

2.3 Exploratory Data Analysis

2.3.1 Correlation Matrix

Correlation matrix shows that interest rate has value of 0.5 positive collinearity with historical default which is the highest correlation. The rest of the feature doesn't correlate much.

2.3.2 Class Imbalance



```
> print(count_ones)    > print(count_zeros)
[1] 5090               [1] 23411
```

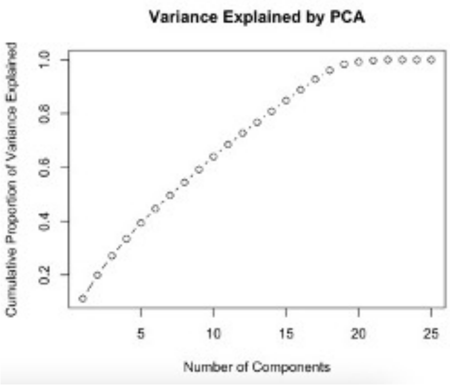
Figure 2.3.2 Imbalance case of defaulting status

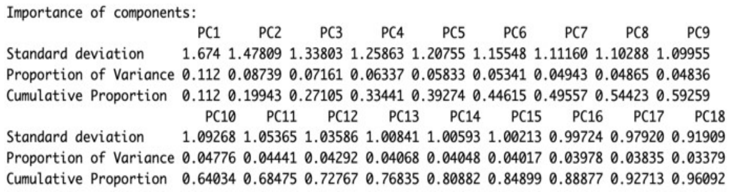
The ratio for default and non-default observations is 5,090 : 23,411 suggesting an imbalance data case. This imbalance case could lead to bias towards the majority class, forming an incorrect understanding of underlying trends in the data. Additionally Nekouei F. (2023), also states that classification algorithm trained on this dataset will unfairly classify minority classes towards majority classes.

2.3.3 Stratified Sampling

Considering classification algorithm will be implemented, it is a must to provide an equal amount of data for each class, unless an intentional weight was given for a specific purpose. Nekouei F. (2023) mentions that it is crucial for both training and test sets have the same proportion through stratification. On top of that, stratification make sure trend or correlations among classes are retained after cross validation.

2.4 Principal Component Analysis





```
Importance of components:
                         PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     1.674 1.47809 1.33803 1.25863 1.20755 1.15548 1.11160 1.10288 1.09955
Proportion of Variance 0.112 0.08739 0.07161 0.06337 0.05833 0.05341 0.04943 0.04865 0.04836
Cumulative Proportion  0.112 0.19943 0.27105 0.33441 0.39274 0.44615 0.49557 0.54423 0.59259
                          PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18
Standard deviation     1.09268 1.05365 1.03586 1.00841 1.00593 1.00213 0.99724 0.97920 0.91909
Proportion of Variance 0.04776 0.04441 0.04292 0.04068 0.04048 0.04017 0.03978 0.03835 0.03379
Cumulative Proportion  0.64034 0.68475 0.72767 0.76835 0.80882 0.84899 0.88877 0.92713 0.96092
```

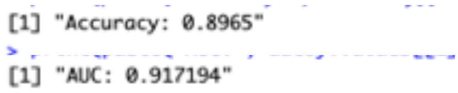
Figure 2.4 Dimensionality reduction and discover underlying pattern

Generally, PCA is used to reduce dimensionality, simplifying an over complex data with a lot of dimension. In this case, PCA will aid in capturing variance since the dataset only have one medium correlated feature which is interest rate, it will be hard for the model to detect underlying pattern with data provided. An IBM (2023) writer says, by identifying principal components that capture most variance, PCA can assist in informing the underlying patterns within the dataset.

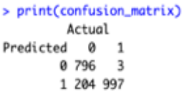
Figure on the left shows is a significant variance explained increase up to PC 18 and after which it hits the plateau. With 18 existing principal component, PCA accumulates 96.09% of variance explained in total. The PCA conducted shows an expected trend, proven by the decreasing standard deviation and proportion of variance explained per component.

2.5 Modelling & Evaluation

2.5.1 Logistic Regression with PCA



```
[1] "Accuracy: 0.8965"
>
[1] "AUC: 0.917194"
```



```
> print(confusion_matrix)
         Actual
Predicted   0   1
        0 796   3
        1 204 997
```

Figure 2.5.1 Confusion matrix

Model has an accuracy of 89.65%, showing 3 false predictions for customers who are currently in default and 204 false predictions for customers who have non-default status. The performance also highly effective at distinguishing those two classes with 91.71% AUC.

2.5.2 Logistic Model Interpretation

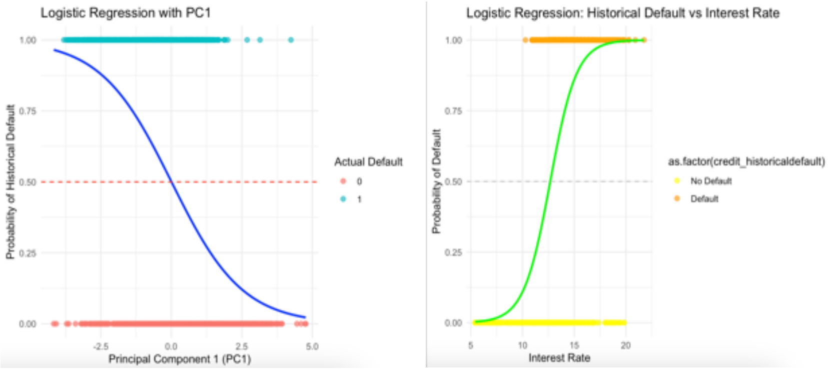
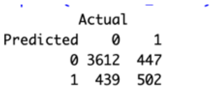


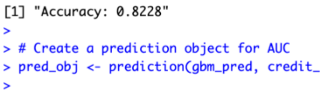
Figure 2.5.2 Model Interpretation by comparing with interest rate

According to Keboola (2022), it is difficult to tell which are the most important features in the dataset after computing principal components, so single interest rate feature will be selected to compare and draw inference. The default rate increases rapidly at interest rate of 7% onwards, showing the risk of loan approval for certain percentage onward. With both figures combined, adding more features aside from interest rate provide more flexibility on loan approval.

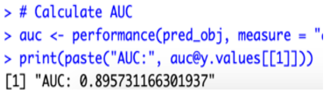
2.5.3 XG Boost



```
         Actual
Predicted    0    1
        0 3612  447
        1  439  502
```



```
[1] "Accuracy: 0.8228"
>
> # Create a prediction object for AUC
> pred_obj <- prediction(gbm_pred, credit_
>
```



```
> # Calculate AUC
> auc <- performance(pred_obj, measure = "
> print(paste("AUC:", auc@y.values[[1]]))
[1] "AUC: 0.895731166301937"
```

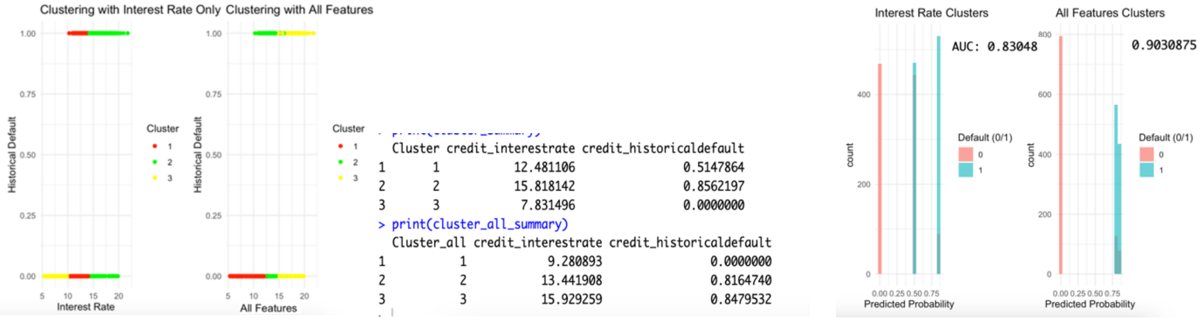
Figure 2.5.3 Comparing logistic model with XG Boost model

Another appropriate algorithm to construct a model for this credit default case is adopting XG Boost Algorithm. Brownlee (2020) stated that XG Boost improves the performance of imbalanced cases. The algorithm also provides a way to tune the training algorithm by adding weighted coefficients for misclassification of the minority class.

Since XG Boost model doesn’t benefit from stratified sampling and PCA, It resulted in less accurate separation with 82.28% of accuracy and AUC 89.57%.

2.5.4 Semi-supervised Learning

Semi-supervised learning may raise the model complexity, a study by Dan et al. (2018) suggested SSL can achieve higher accuracy compared to typical supervised learning depending on the relationship between labelled data. Similar to PCA, it becomes difficult to infer feature contribution after data points are grouped per clustering label. Comparing SSL model with logistic regression model using PCA, SSL model falls short by ~1.5% AUC score.



```
  Cluster credit_interestrate credit_historicaldefault
1       1           12.481106                0.5147864
2       2           15.818142                0.8562197
3       3            7.831496                0.0000000
> print(cluster_all_summary)
  Cluster_all credit_interestrate credit_historicaldefault
1           1            9.280893                0.0000000
2           2           13.441908                0.8164740
3           3           15.929259                0.8479532
```

Figure 2.5.4 Leveraging cluster labels from K-means for classification

With 3 clusters for K-means clustering, model achieves AUC of 83.48% for interest rate feature alone and AUC of 90.30% for all features. However, there is a notable overlapping of clustered data for all feature shown in left figure. This clustering analysis draw two important conclusion: an interest rate of 7.83% or lower ensures no likelihood of default, and by adding other features it could push interest rate threshold further to 9.28%. Furthermore, interest rate from 7.83% to 12.48% also reveal the chance of default at 51.47% while using more feature overlap this findings.

2.6 Summary

Exploring the reason why certain customer default is crucial to maintain financial stability, by leveraging PCA to capture the principal component, the model secure an accuracy of 89.65% and 91.71% AUC. Inferring a very high accuracy of prediction to discriminate both default and non-default customer.

Interest rate as key feature could help indicate the likelihood of defaulting for customer. Since interest rate is derived from a company's regulation to provide the most suitable rate for each customer, the most effective way to understand the complexity of interest rate is by exploring how each regulation provided interact with each other, ultimately determining the best rate. Both clustering and logistic regression with interest rate feature demonstrates a strict and concise, non-overlapping line that clearly set boundaries between classes. Showing the power of interest rate variable in predicting the possibility of defaulting.

Another key takeaway is how SSL shows that interest rate could utilize the other feature to increase its threshold of correctly predicting the non-default customer, in this case from 7.83% alone up to 9.28% by taking advantage of other features considering the need of expertise to verify the credibility of other feature.

References

1. King and Snohomish Counties. (2025). *Property management for investors.* T-Square Properties. Retrieved January 17, 2025, from https://tsquareproperties.net/property-management-for-investors/#:~:text=King%20and%20Snohomish%20Counties%20have,in%20value%20over%20the%20years.
2. Kaushik, S. (n.d.). *House price prediction: A simple guide with scikit-learn and linear regression.* Medium. Retrieved January 17, 2025, from https://medium.com/@kaushiksimran827/house-price-prediction-a-simple-guide-with-scikit-learn-and-linear-regression-f91a27b9d650
3. Alooba. (n.d.). *Handling missing values in data science*. Retrieved from https://www.alooba.com/skills/concepts/data-science/handling-missing-values/
4. Etherington, T. R. (2021). Mahalanobis distances for ecological niche modelling and outlier detection: Implications of sample size, error, and bias for selecting and parameterising a multivariate location and scatter method. *PeerJ*, *9*, e11436. https://doi.org/10.7717/peerj.11436Bruin, E. (2017). *House prices: Lasso, XGBoost, and a detailed EDA*. Kaggle. https://www.kaggle.com/code/erikbruin/house-prices-lasso-xgboost-and-a-detailed-eda
5. Tilii7. (2020, March 25). *Removing collinearity -- 0.5878 with 5 features*. Kaggle. https://www.kaggle.com/code/tilii7/removing-colinearity-0-5878-with-5-features
6. Zhang, H., Guo, X., & Xie, X. (2020). A pitfall of learning from user-generated data: In-depth analysis of subjective class problem. *arXiv*. https://arxiv.org/abs/2003.10621
7. St Nau, R. (n.d.). *Testing the assumptions of linear regression*. Duke University. https://people.duke.edu/~rnau/testing.htm
8. Penn State University. (n.d.). *Lesson 13: Weighted least squares regression*. STAT 501: Regression Methods. https://online.stat.psu.edu/stat501/lesson/13/13.1
9. ProjectPro. (2024, October 28). *8 feature engineering techniques for machine learning*. ProjectPro. https://www.projectpro.io/article/8-feature-engineering-techniques-for-machine-learning/423
10. Analytics Vidhya. (2021, May 19). *The game of increasing R-squared in a regression model*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/the-game-of-increasing-r-squared-in-a-regression-model/
11. Penn State University. (n.d.). *The coefficient of determination, r-squared*. STAT 462: Applied Regression Analysis. https://online.stat.psu.edu/stat462/node/95/

12. University of Alabama at Birmingham. (2021, June 28). *Sample size in machine learning and artificial intelligence*. University of Alabama at Birmingham. https://sites.uab.edu/periop-datascience/2021/06/28/sample-size-in-machine-learning-and-artificial-intelligence/
13. Department of Statistics, Pennsylvania State University. (n.d.). *Lesson 12: Multicollinearity & other regression pitfalls*. In *STAT 501: Regression methods*. Retrieved January 19, 2025, from https://online.stat.psu.edu/stat501/book/export/html/981
14. Zhang, R. (2024). *Feature scaling*. In *UCI Math 10, Fall 2024*. Retrieved January 19, 2025, from https://rayzhangzirui.github.io/math10fa24/notes/feature_scaling.html
15. Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). *Extended comparisons of best subset selection, forward stepwise selection, and the lasso*. arXiv. Retrieved from https://arxiv.org/abs/1707.08692
16. Block Sidge. (2017, January 03). The “median price” is used as the most common indicator for the property market. Block Sidge. https://www.blocksidge.com.au/median-house-price-useful/#:~:text=The%20“median%20price”%20is%20used,consumer%20sentiment%20and%20market%20conditions
17. Akinjole, A., Shobayo, O., Popoola, J., Okoyeigbo, O., & Ogunleye, B. (2024, October 31).. *Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction. MDPI*. https://www.mdpi.com/2227-7390/12/21/3423#:~:text=Predicting%20credit%20default%20risk%20is,thereby%20maintaining%20profitability%20and%20stability
18. Nekouei, F. (2023). *Imbalanced Personal Bank Loan Classification*. *Kaggle.* https://www.kaggle.com/code/farzadnekouei/imbalanced-personal-bank-loan-classification
19. IBM. (2023, December 8). What is principal component analysis *(PCA)? IBM.* https://www.ibm.com/think/topics/principal-component-analysis
20. Keboola. (2022, April 2). *A guide to principal component analysis (PCA) for machine learning. Keboola.* https://www.keboola.com/blog/pca-machine-learning
21. Brownlee, J. (2020, August 21). *How to configure XGBoost for imbalanced classification. Machine Learning Mastery.* https://machinelearningmastery.com/xgboost-for-imbalanced-classification/
22. Dan, C., Liu, L., Aragam, B., Ravikumar, P., & Xing, E. P. (2018). *The Sample Complexity of Semi-Supervised Learning with Nonparametric Mixture Models (NeurIPS 2018), 31, 1–11.* https://arxiv.org/abs/1809.03073