King County House Price, Health Insurance Crossell, Default Of Credit Card Prediction

Nivedita Vishwanath Hiremath
MSc in Data Analytics
x21108471@student.ncirl.ie

Abstract— **In this paper, we have applied various machine learning models on the datasets.For the 'King county house price prediction' Multilinear regression, Ridge regression, Bagging, Random forest, and Gradient boosting are applied. For the 'Health insurance cross-sell' Binary logistic regression, Random forest, and Naivebayes are applied. For the final dataset 'Credit card default' Decision trees, SVM, Xtreme gradient boosting are applied. For each model, cross-validation is done with proper reason.**

I. Introduction

Machine learning is a subfield of artificial intelligence that analyzes the data and makes a decision. It is currently the most advanced solution for a wide range of issues. This motivated to learn the process followed in building machine learning models. We have selected the three datasets for the analysis

A. King county house sale prediction

The dataset is taken from Kaggle[1]. The dataset has 21613 observations and 21 columns. House price prediction will help people to estimate their budget to either buy or sell it. Research questions for this dataset

1. Can we predict house prices by using this data?

2. Are the model evaluation methods producing adequate results?

B. Health insurance cross-sell

The dataset is taken from Kaggle[2]. The dataset has 381109 observations and 12 columns. Insurance cross-sell prediction will help the company to target the audience and use communication strategy to buy vehicle insurance. Research questions for this dataset are-

1. Whether the customer would be interested in vehicle insurance using historical data of health insurance purchased customer?

2. Who are the target customers interested in buying vehicle insurance?

C. Credit card default

The dataset is taken from Kaggle[3]. The dataset has 30000 observations and 25 columns. Identifying and classifying the people to default on credit card loans will help banking in mitigating the financial loss. Research questions for this dataset is –

Using previous payment history, how accurately can we predict whether the customer will pay the credit card bill for the following month?

II. Related Work

Several articles are published for the following datasets.

For the king county dataset, In this paper [1] house resale prediction is done using Logistic regression, Decision tree, Naïve Bayes, Random forest, Ada boosting, and here they considered accuracy as the performance metrics to decide the best algorithm Adaboost and Decision tree C5.0 emerged as winners with the highest accuracy. The drawback of this work is they did not include the cross-validation methods before building the models. In this paper [2] authors have applied Linear models, Tree-based models , Deep learning model, and Catboost and compared the results with RMSE the Catboost was showing the least RMSE and it was used for the prediction in testing data. This is another way of selecting the best model. In this paper [3] authors have used the King County dataset to evaluate the predictive performances of popular statistical learning methods. Linear Regression, Elastic Net, Random Forest, Gradient Boosting Method, Support Vector Machines, and Neural Network are the methods used. Although the model was cross-validated but they were limited the data of 10% to train the models, so in future, the model may show the variance in predicting results. The authors in this paper [4] evaluate Stochastic Gradient Descent, Stochastic Dual Gradient Ascent, Gradient Tree Boosting and clustering using K-Means to predict real-estate prices. When all models were compared, the Gradient tree boosting model produced the best results, with the highest R2 and the lowest

[1] https://www.kaggle.com/harlfoxem/housesalesprediction

[2] https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction

[3] https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

normalized root mean square error. Authors in this paper [5] initially figured out the features required in prediction using PCA, Information value computation information and VIF after the feature selection they have applied ANN,SVM and Random forest algorithm although Random forest performed better they chose SVM because Random forest prone to overfit the data. The advantage of this work is they have applied dimensionality reduction techniques to select feature variables before applying to the train model we can consider this in our model building process.

For the Health insurance cross-sell, The goal of this work[6] was to create a model that could anticipate policyholders' reactions to automobile insurance. Insurance businesses can utilize this to plan how to reach out to their clients and maximize their business model and income. The aim here is to maximize recall, so that insurance companies may later send promotions to all potential clients. There were three machine learning algorithms used: Logistic Regression, Decision Tree, and Random Forest. Logistic regression performed better with respective recall/sensitivity values. The drawback of this work is they did not include cross validation methods before building the models.

For the Credit card default , The aim of this paper [7] is to apply SVM algorithm in prediction of credit card default and to enhance the accuracy of the SVM they have used LS-SVM and ensemble method.LS-SVM has performed better with an accuracy. The disadvantage of this project would be they just considered accuracy as a criteria to choose best algorithm instead they would have also included sensitivity and specificity to decide the algorithm, because models ability to predict defaulters is more important. The authors of this paper [8] proposed a method to handle imbalance data called 'SRIPPER' it was compared with 'SMOTE' and 'RIPPER' the 'SRIPPER' Showed better accuracy drawback of this work is they applied above technique only to a few algorithms namely SVM, BP, C4.5.The author of this paper [9] identified the imbalance in the data and created a new weighted model to improve the prediction of default class, the weights of default and non-default as 1:3:5, and applied machine learning algorithms namely Logistic, Decision trees, Ada boosting and Random forest. The authors of this paper[10] says that even after handling imbalanced data the models tend to bias towards majority class samples so they proposed a new solution called Multiple classifiers system(MCS) this detects anomalies in those minorities in the dataset. But the drawback of this work is that author did not show proper results comparing other sampling techniques called SMOTE, Udersampling or Oversampling. In another paper[11] the authors have applied eight models namely Logistic regression, Neural network, radial basis function neural network,SVM,case base reasoning, and decision trees cross validation was also applied to the models but the limitation of paper was on chosen dataset ,the dataset has only 1000 records. In another paper [12] authors analysing the credit risk for the small and medium-sized enterprises, here the study is done applying logistic regression and resampling the data ,but drawback is it has less feature variables to detect default.

III. Methodology

In this project, we followed KDD (Knowledge discovery in databases) method to extract the knowledge from the data[4].This involves several steps-

a. Data selection

b. Data cleaning and pre-processing

c. Feature selection and projection

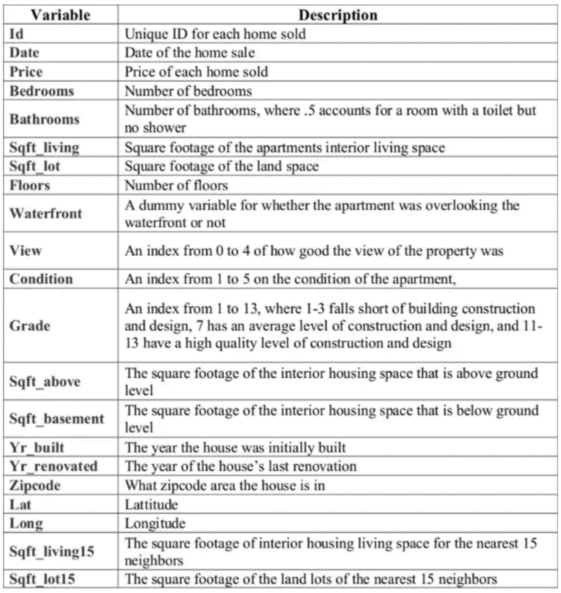
d. Choosing data mining algorithms

e. Interpreting mined results

A. King county house price sale

a. **Dataset selection** -The dataset includes house sale prices between May 2014 and May 2015 of King County, USA. The target variable is 'price'

Dataset Dimensions – 21613 rows,21 columns



Variable	Description
Id	Unique ID for each home sold
Date	Date of the home sale
Price	Price of each home sold
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
Sqft_living	Square footage of the apartments interior living space
Sqft_lot	Square footage of the land space
Floors	Number of floors
Waterfront	A dummy variable for whether the apartment was overlooking the waterfront or not
View	An index from 0 to 4 of how good the view of the property was
Condition	An index from 1 to 5 on the condition of the apartment,
Grade	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design
Sqft_above	The square footage of the interior housing space that is above ground level
Sqft_basement	The square footage of the interior housing space that is below ground level
Yr_built	The year the house was initially built
Yr_renovated	The year of the house's last renovation
Zipcode	What zipcode area the house is in
Lat	Lattitude
Long	Longitude
Sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors
Sqft_lot15	The square footage of the land lots of the nearest 15 neighbors

Fig. 1. King county house price dataset description

[4] http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

b. **Data cleaning and preprocessing** - The first step in data cleaning is checking for null values. The dataset selected here does not contain any null values

c. **Feature selection and projection** -
Fig. 2 shows bathrooms, sqft_living, grade, sqft_above, sqft_living15 highest correlation with price.

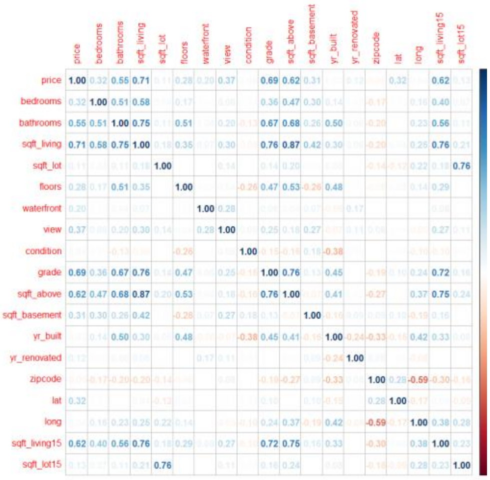


Fig. 2. Correlation matrix king county house price dataset

Boxplots of categorical variables like bedrooms, bathrooms, zipcode, view, and grade were drawn against price, the house price value seemed to increase with the increase in price. But one interesting anomaly was found between bedroom and price Fig. 3 bedrooms with 33 seem to be an outlier, by investigating that row we found that it has 33 bedrooms and 1 floor which is not feasible ideally so removing that row from our dataset.

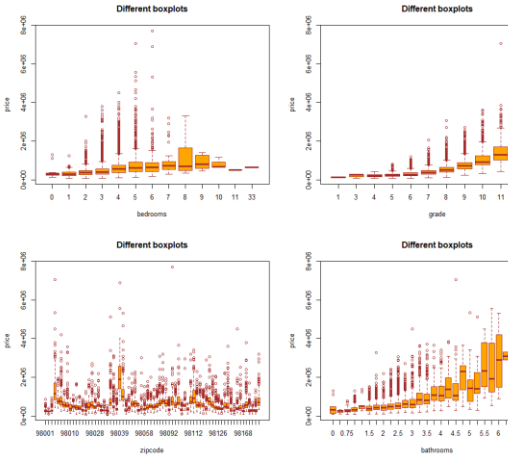


Fig. 3. Boxplots of categorical variables

d. **Choosing data mining algorithms** - By analyzing the dataset and correlation matrix, we assumed that the id and date column does not give much information in analyzing house price hence it was removed in the model building process. Dataset was divided into train and test subset in 70:30 ratio. For this dataset, we decided to apply Multilinear regression, Ridge regression, Regression tree, Bagging(bootstrap aggregation), Random forest, Gradient boosting.Cross-validation hyperparameter tuning was done for the algorithms and the best algorithm was chosen based on the least MSE(mean square error value).

Multilinear regression- We used trial and error of taking feature variables steps we have done to get a good Adjusted-R squared value. The model with a good Adjusted-R squared value was selected. To prove it as the best model, we further went to satisfy Gauss Markova - Linearity, Homoscedasticity, Multicollinearity, and Influential Observation. The model selected failed to prove Homoscedasticity. To rectify this error another model was built by applying log transformation to dependent variable price. It gave a good Adjusted R- squared value of 0.7665 and a residual standard error of 0.2557.The Fig. 4. Shows diagnostic plots for this model.

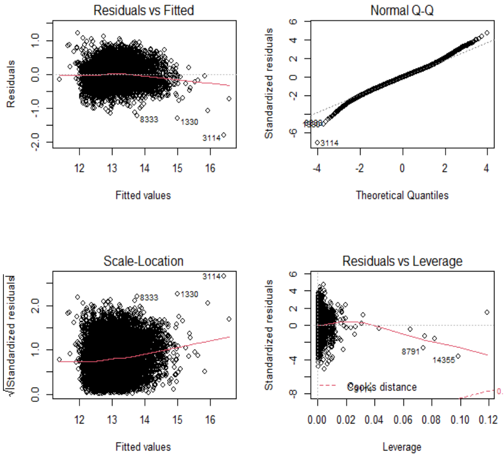


Fig. 4. Diagnostic plot for model

Proving assumption satisfaction is done by conducting test. ncvTest for Homoscedasticity, to our model even after applying log to response variable test failed need to think of another form of transformation, durbinWatsonTest was used for Autocorrelation between errors we got D-W statistic value nearer 1.99283 nearer value to 2 so it was satisfied, VIF for the absence of multicollinearity this test was satisfied as all the features had a value less than 10 and lastly cooks distance was calculated and all values remained lesser than 1, hence proved no influential data points. To this model, the prediction was done on test data. It showed linear relation with predicted and actual value Fig. 5. and MSE(mean square error) was calculated 0.06465.

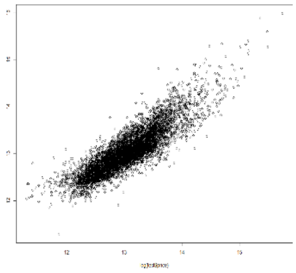


Fig. 5. Predicted and actual value

Ridge regression - Ridge regression introduces a little amount of bias as penalty Λ value, resulting in a decrease in variance[5] it has the advantage over simple multilinear regression in that it can improve predictions produced from new data (i.e. lower variance) by making the predictions less susceptible to the training data when sample sizes are limited. Cross-validation was used to determine the model's flexibility, 1 standard error away from the lambda(Λ) that provides the lowest mean square error(MSE). Fig 6. shows with an increase in lambda MSE also increases. The chosen lambda value is 0.05942248. Prediction on test data was done using lambda value obtained MSE value 0.06391 was noted.

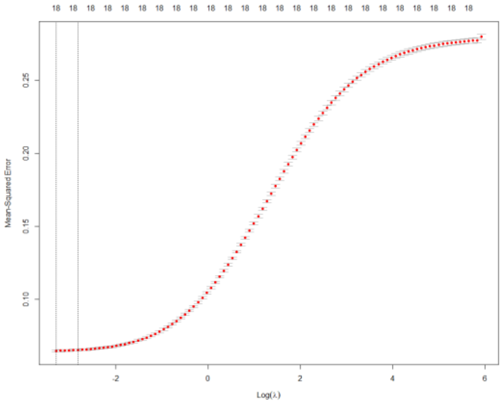


Fig. 6. Lambda vs MSE

Regression tree - The data is better reflected by trees than by a straight line. Each leaf node in a tree corresponds to a particular cluster of observations' average value. The regression tree is applied to the training dataset. Fig. 7. shows the tree with 8 terminal nodes.

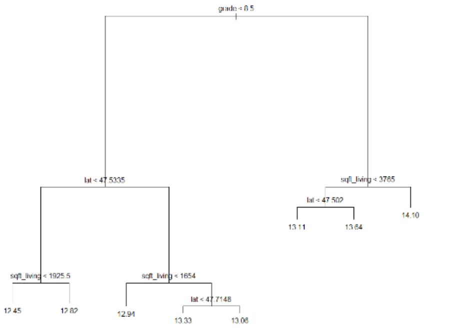


Fig. 7. Regression tree

To check the model’s performance regression tree was applied to the testing set. The cross-validation was conducted to check whether we need to prune the tree, the pruning process was guided by the deviance error rate. Fig. 8. shows the plot of the cross-validation result, we can see the lowest deviance error rate in the tree was with 8 terminal nodes. Thus we need to stick with our original tree.

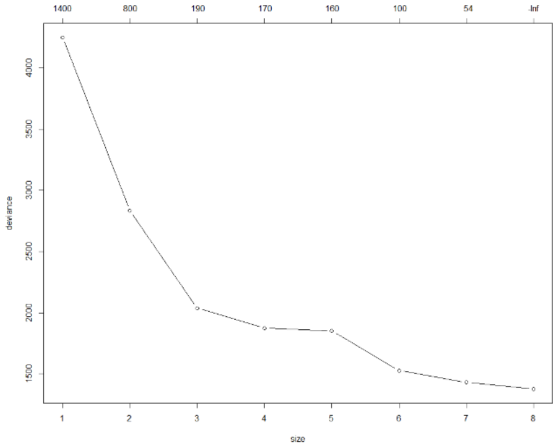


Fig. 8. Deviance error rate with tree nodes

The MSE was also calculated. The MSE value obtained was 0.09155.

Bagging - It is applied to the binary trees, at each node it searches for all the features that best splits the data at that node. We are using node size as 8 because we found that as best in tree regression cross-validation. To this MSE received was 0.08615 with 69.33% variation explanation. To assess the model’s accuracy, prediction of the prices was

[5] . https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

done on the test dataset and it was plotted against the actual price on the test subset. Fig. 9. Shows residuals following standard distribution

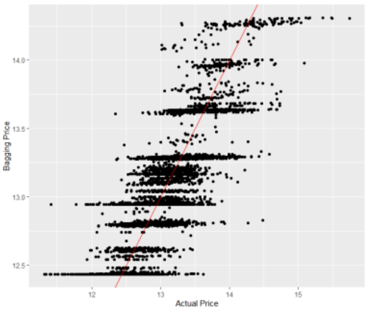


Fig. 9. Actual Price vs Bagging price

Random Forest **-** Initially, random forest was run across all the 19 predictors taken as features at each split. The proportion samples incorrectly predicted known as the Out-of-Bag error rate is calculated for features in each split. Fig. 10. shows MSE was low from the 8th split Hence tuned our model by taking 8 predictors with an increased number of trees.

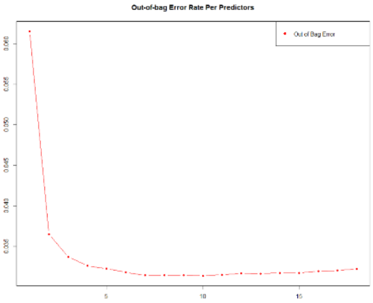


Fig. 10. MSE Vs features at each split

To assess the model quality it was run against the test data. Fig. 11. is a plot of predicted price with test price. The MSE obtained from this model was 0.03829. The disadvantage of Random forest is it took 30 minutes to run the model.

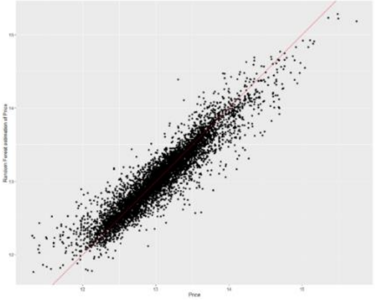


Fig. 11. Actual price vs Random forest predict price

Boosting Regression – To deal with the problem of lower bias and high variance. It uses something called learning rate to scale contribution from the new tree. The learning rate is a value between 0 and 1. Here we had initially set the hyperparameter grid of learning rate to 0.01,0.1,0.3 and checked which will yield the least MSE values. By observing the lower MSE values, the parameters are narrowed to 0.01,0.05,0.1.The final MSE value obtained was 0.02683.Plot was drawn against price predicted vs actual price. Fig. 12.

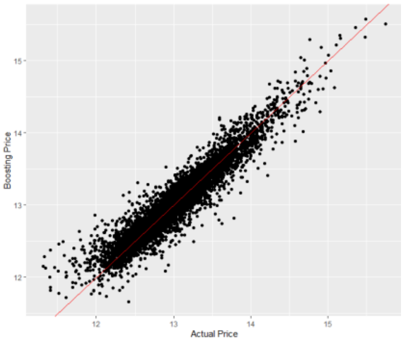


Fig. 12 Price vs. Boosting Predicted Price

e. **Interpreting mined results** –

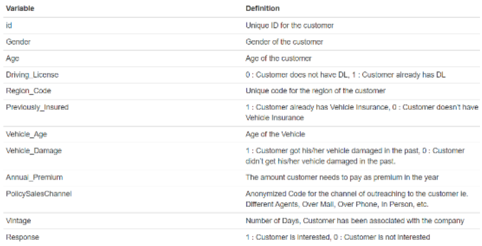
TABLE I Different algorithms with MSE values

Algorithm	***MSE***
Multilinear Regression	0.06465
Ridge Regression	0.06391
Regression Tree	0.09155
Bagging	0.08615
Random Forest	0.03829
Gradient Boosting	0.02683

Gradient boosting performed best with the least MSE value of 0.02683.

B. Health Insurance Cross-sell

a. **Dataset selection** -The dataset includes the customer's past details of the health insurance company. The company is interested in knowing target customers who might be interested in buying vehicle insurance. The target variable is 'Response'.
Dataset Description and dimension – 381109 rows and 12 columns



Variable	Definition
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving_License	0 : Customer does not have DL, 1 : Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
Vehicle_Age	Age of the Vehicle
Vehicle_Damage	1 : Customer got his/her vehicle damaged in the past, 0 : Customer didn't get his/her vehicle damaged in the past.
Annual_Premium	The amount customer needs to pay as premium in the year
PolicySalesChannel	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
Vintage	Number of Days, Customer has been associated with the company
Response	1 : Customer is interested, 0 : Customer is not interested

Fig. 13. Dataset description of health insurance cross-sell

b. **Data cleaning and preprocessing**- No null values found in the dataset. From above Fig. 13. we can say that Gender, Vehicle_Age, Vehicle_Damage are non-numeric values. Label encoder is used to convert those to numeric.

c. **Feature selection and projection** - The feature variables such as Age, Previously_insured, gender,driving license are more affecting the response variable. These can be verified by plotting plots against response variables and plotting correlation matrix.

Fig. 14 .shows male has majority tend to buy insurance

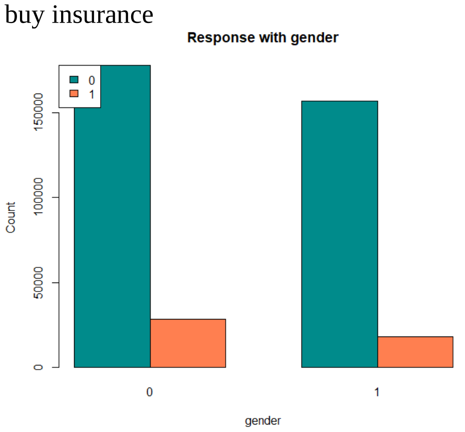


Fig. 14. Response count vs gender

Fig. 15. shows customers with having driving license tend to buy insurance

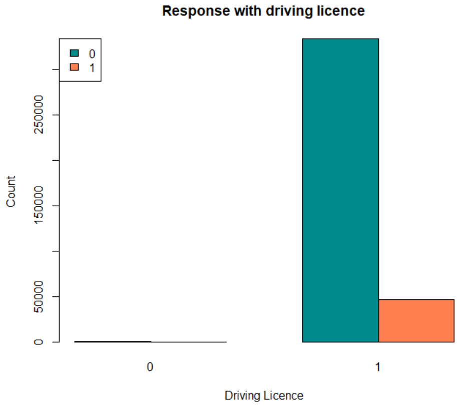


Fig. 15. Response count vs Driving Licence

Fig. 16. shows people tend to buy insurance if they do not have previous insurance

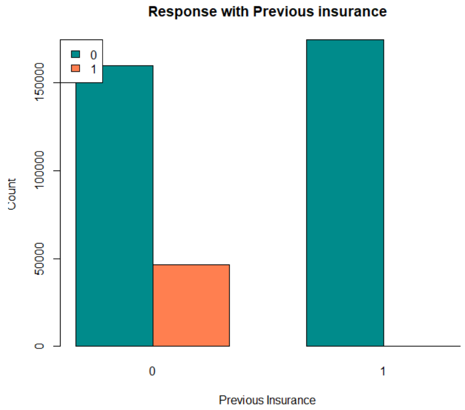


Fig. 16. Response count vs Previous Insurance

Fig. 17. People aged between 30-60 are more interested in buying insurance

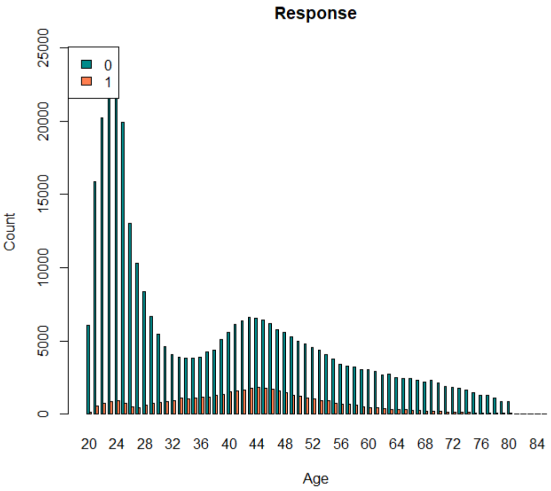


Fig. 17. Response count vs Age

Fig. 18. shows target variable 'Response' is an imbalance problem as the Response variable with the value 1 is significantly lower than the value 0. When we plotted the proportion table it shows 0 as 87% and 1 as 12%.

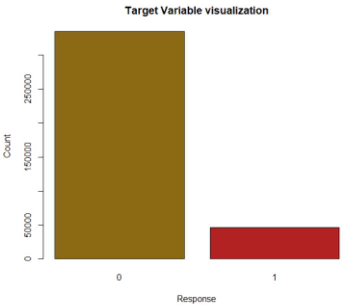


Fig. 18. Response count

d. **Choosing datamining algorithms** - We have used Logistic Regression, Random Forest, and Naïve Bayes. The dataset was divided into training and testing with a

70:30 ratio. As we observed an imbalance in the target variable we have applied SMOTE(Synthetic Minority Over Sampling) technique to balance it. The application of the algorithm is carried twice before and after applying SMOTE.

Logistic regression – Initially when logistic regression was applied to all the feature variables and was found that features 'id',' vintage',' region code' were insignificant and removed in the next iteration. This new model was run against the test set, the prediction was done and prediction probability was plotted in a histogram. By looking at the histogram. There was no probability prediction greater than 0.5 so, the confusion matrix was plotted with the classification threshold greater than 0.25 as a 'Yes' response. ROC-AUC graph was plotted.

Random Forest - Random Forest was applied to the same set of feature variables with a number of trees as 100 and an error rate versus trees plot was drawn to get tress with the least error rate. It was found that chosen number of trees gave the least error rate Fig. 19. Confusion matrix and ROC-AUC curve were drawn to it.

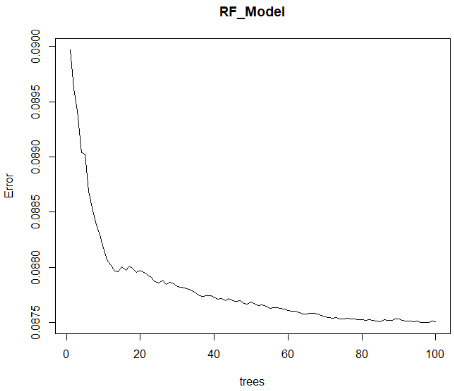


Fig. 19. Error rate vs trees

Naïve Bayes was applied to the same set of feature variables, and the model was applied to the test set. The confusion matrix and ROC-AUC curve were plotted.

After applying SMOTE to the target variable. All the algorithms were re-run. But this time probability classification threshold greater than 05 was chosen as a 'Yes' response. Below are the plots for the ROC-AUC curve after applying SMOTE.

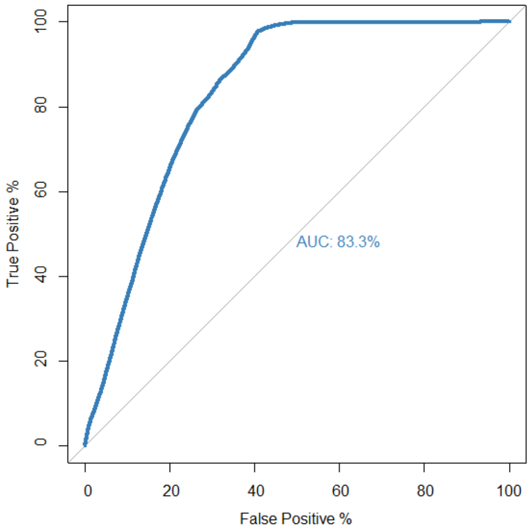


Fig. 20. Logistic ROC-AUC after SMOTE

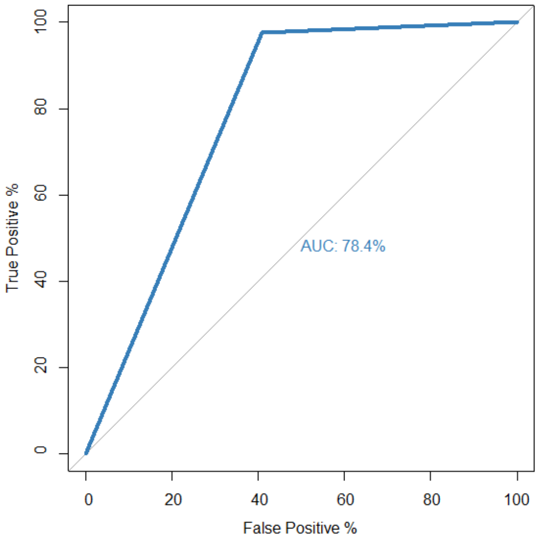


Fig. 21. Random Forest ROC-AUC after SMOTE

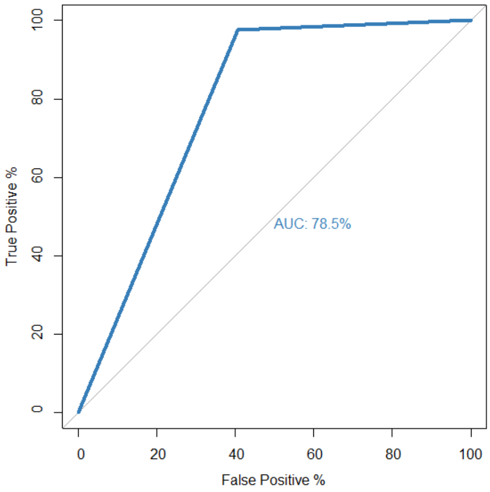


Fig. 22.Naive Bayes ROC-AUC after SMOTE

e. Interpreting mined results-

TABLE II Different algorithms with accuracy and ROC-AUC values before SMOTE

Algorithms	*Accuracy*	*ROC-AUC*
Logistic	79%	83.2%
Random Forest	78%	85.5%
Naïve Bayes	63%	78.4%

TABLE III Different algorithms with accuracy and ROC-AUC values after SMOTE

Algorithms	*Accuracy*	*ROC-AUC*
Logistic	78%	83.3%
Random Forest	79%	79.7%
Naïve Bayes	78%	78.5%

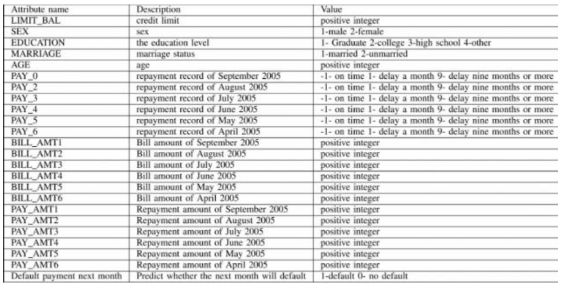
We have considered the Accuracy and ROC-AUC factor to declare the best model because ROC-AUC helps in analyzing the distinction between the classes[6] as 'yes' or 'no' in our case. Higher the curve grater the ability to distinguish. By this Health insurance company can target the customers to attract them to buy vehicle insurance by giving offers and notifications.

By looking at TABLE II and TABLE III. Binary logistic and Random forest does a better job. It makes sense to choose Binary logistic as a winner because it has a higher ROC-AUC curve and it took less time than Random forest to run.

By looking at the above diagnostic visualizations we can say that target customers are aged between 30-60, customers with having driving licenses tend to buy insurance and Male has majority tend to buy insurance than females.

C. Credit card default

a. **Data selection** – This dataset contains information on credit card clients from Taiwan from April 2005 to September 2005.The target variable is 'default.payment.next.month'.
Dataset dimension and description – 30000 and 25 columns



Attribute name	Description	Value
LIMIT_BAL	credit limit	positive integer
SEX	sex	1-male 2-female
EDUCATION	the education level	1- Graduate 2-college 3-high school 4-other
MARRIAGE	marriage status	1-married 2-unmarried
AGE	age	positive integer
PAY_0	repayment record of September 2005	-1- on time 1- delay a month 9- delay nine months or more
PAY_2	repayment record of August 2005	-1- on time 1- delay a month 9- delay nine months or more
PAY_3	repayment record of July 2005	-1- on time 1- delay a month 9- delay nine months or more
PAY_4	repayment record of June 2005	-1- on time 1- delay a month 9- delay nine months or more
PAY_5	repayment record of May 2005	-1- on time 1- delay a month 9- delay nine months or more
PAY_6	repayment record of April 2005	-1- on time 1- delay a month 9- delay nine months or more
BILL_AMT1	Bill amount of September 2005	positive integer
BILL_AMT2	Bill amount of August 2005	positive integer
BILL_AMT3	Bill amount of July 2005	positive integer
BILL_AMT4	Bill amount of June 2005	positive integer
BILL_AMT5	Bill amount of May 2005	positive integer
BILL_AMT6	Bill amount of April 2005	positive integer
PAY_AMT1	Repayment amount of September 2005	positive integer
PAY_AMT2	Repayment amount of August 2005	positive integer
PAY_AMT3	Repayment amount of July 2005	positive integer
PAY_AMT4	Repayment amount of June 2005	positive integer
PAY_AMT5	Repayment amount of May 2005	positive integer
PAY_AMT6	Repayment amount of April 2005	positive integer
Default payment next month	Predict whether the next month will default	1-default 0- no default

Fig. 23. Dataset description of credit default

b. **Data cleaning and preprocessing** - Missing values were in a dataset and found no missing values. By plotting boxplot Fig. 24. We found there are many outliers.

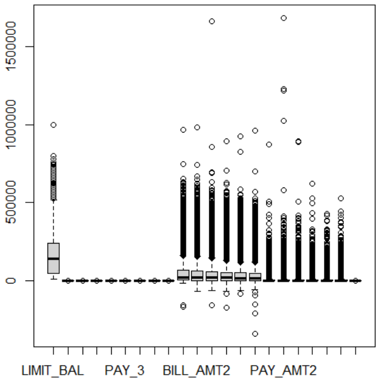


Fig. 24. Outliers from dataset

Removed extreme outliers from the dataset. Bar graph was plotted for target variable Fig. 25 and it was found that the percentage of default class is 22%, so the data imbalance is not significant. By plotting the correlation matrix we found 'id',' sex', and 'marriage' are not much significant in predicting the target variable so those feature variables were removed in model building steps. Dataset was divided into training and testing in an 80:20 ratio.

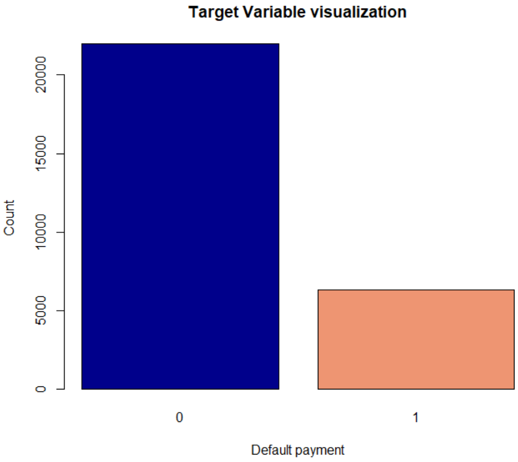


Fig. 25. Count of default.payment.next.month

[6] https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Fig. 26 shows female has more tendency to default than male

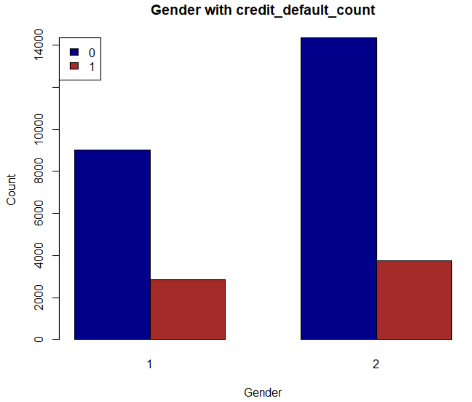


Fig. 26. Gender vs Default count

Fig. 27. Shows customer with lesser credit limit are more to default ,especially lesser than 50000

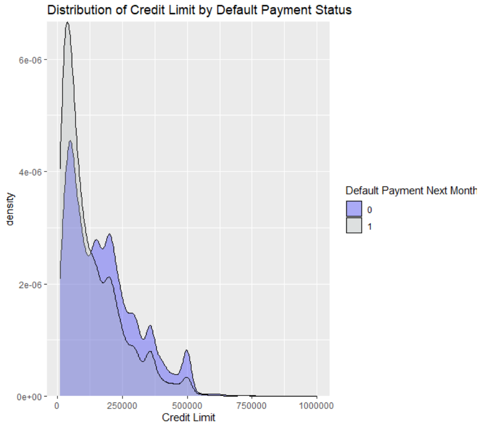


Fig. 27. Creditlimit vs Default count

c. **Choosing datamining algorithm** - For this dataset we chose to apply Decision tree, SVM and Xtreme gradient boosting. The experiment was conducted twice by shuffling the train and test data,this ensures every item has a chance to appear in any position[7]

For the Decision tree we have used library(C50) it uses information entropy to best split the data at the node by minimizing the computed entropy value into purer classes[8].

Training and testing the model was done to a Decision tree, SVM, and Xtreme gradient boosting. Here is one thing to note: SVM took 20minutes to run the training model.

d. **Interpreting the mined results** - The Confusion matrix was noted after applying all the models to test data. Below is the table which describes the results of both iterations.

TABLE III First iteration with the dataset

Algorithm	*Accuracy*	*Sensitivity*	*Specificity*
SVM	75%	81%	68%
Decision Tree	82%	83%	68%
Xtreme Gradient Boosting	81%	84%	64%

TABLE IV Second iteration with the dataset

Algorithm	*Accuracy*	*Sensitivity*	*Specificity*
SVM	76%	81%	44%
Decision Tree	82%	84%	68%
Xtreme Gradient Boosting	81%	84%	66%

As credit card default is a crucial factor for any bank as they don't want to lose money. The sensitivity factor plays an important role, since sensitivity has the ability to correctly identify the customers with default credit. By looking at the above table Decision Tree and Xtreme Gradient boosting maintained the same consistency in getting Sensitivity, we took Specificity into account as it helps in identifying the customers without default. So by considering these factors Decision Tree wins the race.

IV. Conclusion and Future Work

Overall we have undergone applying several machine learning algorithms. We have handled class imbalance problem , cross-validation was done and also applied hyperparameter tuning techniques to increase model prediction power.

Dataset1: King County House sale prediction: For this dataset Gradient boosting performed best with least MSE value of 0.02683. By plotting the Homogeneity of variance graph we still see some curve-like structure so further analysis should be done to rectify this issue and also 'date'- it represents when the house was sold feature usage, and interaction between feature variables usage might improve the prediction rate.

[7] https://medium.com/100-days-of-algorithms/day-43-shuffle-b5abe4644c23#:~:text=In%20machine%20learning%20we%20often%20need%20to%20shuffle%20data

[8]http://mercury.webster.edu/aleshunas/R_learning_infrastructure/Classification%20of%20data%20using%20decision%20tree%20and%20regression%20tree%20methods.html

Dataset2:Health Insurance Cross-sell: For this dataset. Binary logistic performed better by considering the ROC-AUC graph and accuracy. For further enhancements, we suggest taking different thresholds to predict the model and also using different sampling methods like undersampling, oversampling and a combination of both to get the best model.

Dataset 3: Credit card Default: For this dataset. The decision tree wins the race by getting Sensitivity and Specificity results. For further improvements, we suggest using Dimensional Reduction Techniques and then comparing the model because the dataset has nearly 20 features to predict with one target variable.

V. References

[1] P. Durganjali and M. V. Pujitha, "House Resale Price Prediction Using Classification Algorithms," *IEEE Xplore*, Mar. 01, 2019. https://ieeexplore.ieee.org/abstract/document/8882842 (accessed Sep. 23, 2021).

[2] T. Wang, Y. Wang, and M. Liu, "A Price Prediction Method Based on CatBoost," *IEEE Xplore*, Nov. 01, 2021. https://ieeexplore.ieee.org/document/9637683 (accessed Dec. 24, 2021).

[3] C. Yunjiao, F. Zhuolun, Z. Yuzhe, H. Yilin, and D. Shanshan, "Comparison of Statistical Learning and Predictive Models on Breast Cancer Data and King County Housing Data," Sep. 01, 2017.

[4] B. Chandramouli, "Real-estate price prediction for King County region." Accessed: Dec. 24, 2021. [Online]. Available: http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26646708.pdf.

[5] D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques," *IEEE Xplore*, Sep. 01, 2017. https://ieeexplore.ieee.org/abstract/document/8392275/.

[6] A. Hung, "Health Insurance Cross Sell Prediction," *Guten Tag! I am Amy!*, Dec. 12, 2020. https://gutentagworld.wordpress.com/2020/12/13/health-insurance-cross-sell-prediction/ (accessed Dec. 24, 2021).

[7] A. Lawi and F. Aziz, "Classification of Credit Card Default Clients Using LS-SVM Ensemble," *IEEE Xplore*, Oct. 01, 2018. https://ieeexplore.ieee.org/document/8780427 (accessed Dec. 24, 2021).

[8] P. Xu, Z. Ding, and M. Pan, "An improved credit card users default prediction model based on RIPPER," *IEEE Xplore*, Jul. 01, 2017. https://ieeexplore.ieee.org/document/8393037 (accessed Dec. 24, 2021).

[9] Y. Yu, "The Application of Machine Learning Algorithms in Credit Card Default Prediction," *IEEE Xplore*, Aug. 01, 2020. https://ieeexplore.ieee.org/document/9275986 (accessed Dec. 24, 2021).

[10] S. N. Kalid, K.-H. Ng, G.-K. Tong, and K.-C. Khor, "A Multiple Classifiers System for Anomaly Detection in Credit Card Data With Unbalanced and Overlapped Classes," *IEEE Access*, vol. 8, pp. 28210–28221, 2020, doi: 10.1109/ACCESS.2020.2972009.

[11] J. Zurada, "Could Decision Trees Improve the Classification Accuracy and Interpretability of Loan Granting Decisions?," *IEEE Xplore*, Jan. 01, 2010. https://ieeexplore.ieee.org/document/5428636 (accessed Dec. 24, 2021).

[12] R. Yang, X. Zhou, and W. Wang, "Is the Small and Medium-Sized Enterprises' Credit Default Behavior Affected by Their Owners' Credit Features?," *IEEE Xplore*, Aug. 01, 2011. https://ieeexplore.ieee.org/document/5998460 (accessed Dec. 24, 2021).