BAN 620

CASE 3

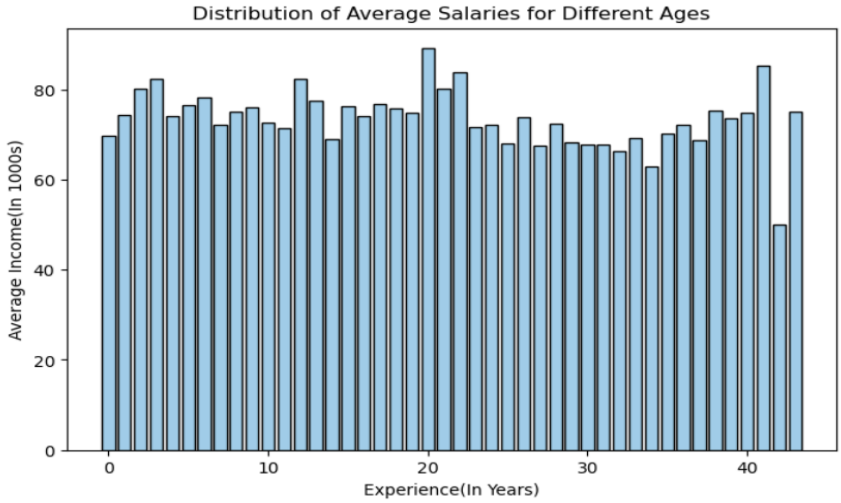
GROUP 6

The given dataset is regarding the Universal bank customer data which has 14 variables, out of which the variable 'Personal Loan' is the variable of interest which has to be predicted as if the customer takes a personal loan or not.

We have removed the variables 'ID', 'Zipcode' which doesn't have any significance on the model. After dropping these variables, we have 5000 records and 12 variables. Since all the variables are numerical there is no need for encoding the data.

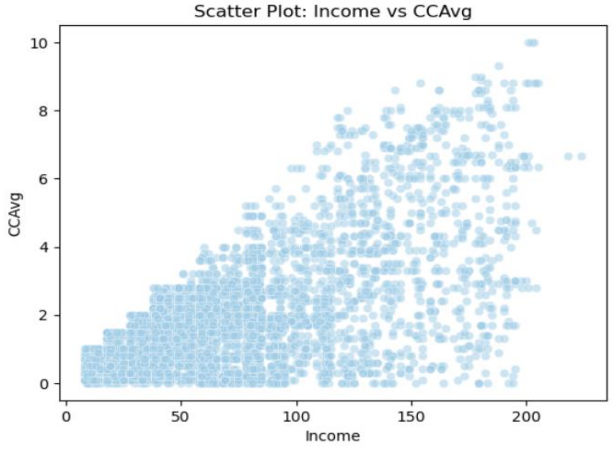
From the dataset, it is observed that there are 52 records identified as outliers(Experience < 0), which is approximately 1% of the data, so these outliers has been removed from data and rest of the data was considered.

Upon doing the basic statistical analysis, it was found that the average age of the customers is 45 years with an average of $73.8K income per year and atleast half of them has a minimum of 20 years of experience.



Among the customers, ~30% of the account holders have credit card issued by the universal bank and they are spending an average of approximately $2000 on their credit card. Among the customers only 10% of the customers hold securities account, whereas only 6% hold CD account with the bank

Also it is proved that the customers income is directly related to their credit card bill amount and shares a positive co-relation.



The main purpose of this model is to predict if a customer will take a personal loan or not. To build this model, data has been classified as training data and validation data, it has been made to 70% and 30% respectively.

To predict the outcome of the model, we have given an unknown record to validate whether this customer will take a loan or not. So, this customer we are giving as new input is an 49 years aged person with 25 years of experience and having an Annual income of $170K, who has an PhD level of education with no mortgage. This person family consists of 3 persons with an average credit card bill of $7600 per month. This person doesn't hold a securities account with bank but hold an CD account utilizing the internet banking facilities.

	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Securities Account	CD Account	Online	CreditCard
0	49	25	170	3	7.6	3	0	0	1	1	1

A KNN model has been developed by standardizing the values in the data set and the new data point was also standardized and fed to the model to predict the outcome. Model has predicted that this person **does not take a Personal loan.** This prediction has been made by considering the outcome of 3 (k=3) nearest neighbors to this new data record.

```
[0.]
Distances [[1.55471989 1.6247933  1.69113469]]
Indices [[1413 1385 1956]]
          zAge  zExperience   zIncome   zFamily    zCCAvg  zEducation  \
2026 -0.222241    -0.379510  1.898524  0.519864  3.535331    0.126672
1985  0.130979     0.238181  2.093161 -0.350092  2.665660    0.126672
2806 -0.045631     0.061698  1.292988  1.389821  2.201835    1.315813

      zMortgage  zSecurities Account  zCD Account   zOnline  zCreditCard  \
2026  -0.564828            -0.336851     3.987175  0.819151     1.567509
1985  -0.564828            -0.336851     3.987175  0.819151     1.567509
2806  -0.564828            -0.336851     3.987175  0.819151     1.567509

      Personal Loan
2026            0.0
1985            0.0
2806            0.0
```

From the above output, it is predicted that personal loan outcome is 0, which means that this new customer will not take the loan. This prediction was made by considering 3 nearest data points which are 1.55,1.62,1.69 Euclidean distance points away from the data point.

This prediction is solely based on the k value of 3, which is not a accurate value, in order to predict more accurately we should consider a particular number of data points (K values) around the prediction data point to get more accurate results.

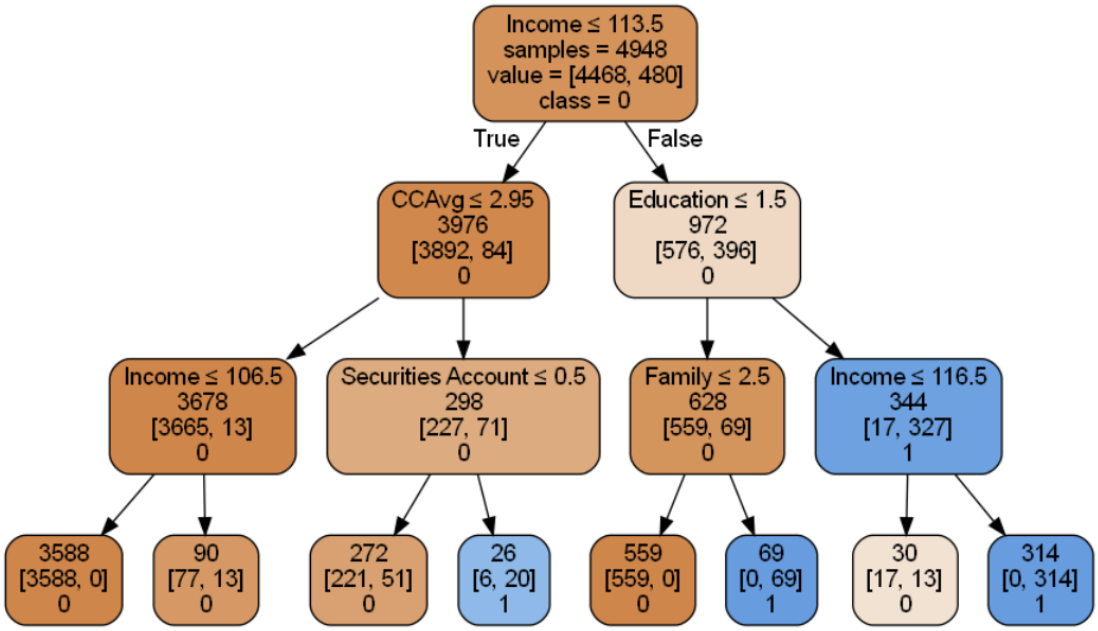
We have considered multiple K values and tested it on the validation data to check for how many values of k, we get highly accurate model. We considered k values from 1 to 25 and cross-checked the accuracy between these numbers and concluded that by considering 12 neighbors around the target data point will give more accurate results i.e., **k=12**

	k	accuracy
0	1	0.829352
1	2	0.891468
2	3	0.878498
3	4	0.894198
4	5	0.892150
5	6	0.896246
6	7	0.894198
7	8	0.896928
8	9	0.896928
9	10	0.896928
10	11	0.896928
11	12	0.897611
12	13	0.896928
13	14	0.897611
14	15	0.897611
15	16	0.897611
16	17	0.897611
17	18	0.897611
18	19	0.897611
19	20	0.897611
20	21	0.897611
21	22	0.897611
22	23	0.897611
23	24	0.897611

At k=12, our model has an accuracy of 89.76%, this is the best possible value of k instead of considering any higher values or lower values as, higher k values cause underfitting and lower/smaller k values causes overfitting. These cases make the model to predict inaccurate by considering the noise and not capturing the patterns in data which is essential to make a prediction.

Along with the KNN model, a classification tree was also developed by considering the decision tree classifier. In this classification tree, all the 11 features were considered and a classification tree was developed, which gives a better understanding of whether a person takes a personal loan or not based on his/her inputs for the variables.

Out[71]:



By looking into the above tree, a rule can be made to check if a person takes this loan or not. One such rule that can be derived from this tree was " If the customer annual income is greater than $113,500 per year and if he/she has an education level of Graduate or more and if that cx income is more than $116,500 then there are high chances that customer will accept the personal loan from bank"

Between the KNN model and classification model, the classification model can be used to predict whether the customer accepts the personal loan or not. This decision was based on the accuracy of each model.

For KNN, the maximum accuracy at k=12 is 89.76$

Whereas, for classification tree the accuracy based on confusion matrix is 98.32%

By considering these accuracies and ease of usage, classification tree is better for predicting the customer acceptance of personal loan offered by the universal bank.

E-Bay Auction Case

Given data is related to the e-bay auctions for multiple categories and their bid prices.

This dataset has 8 variables(`Category, currency, sellerRating, Duration, endDay,ClosePrice, OpenPrice, Competitive, Price_Margin`), out of which we have only 1 variable of interest i.e., competitive. This variable allows us to understand whether the bid for a particular product is competitive or not. This was considered based on the number of bids for that product.

Upon conducting initial statistical analysis on the dataset, it is found that there are 18 categories in which the products are arranged and there is a high variance(std=5973) in the seller rating for each product. The open price of the bid varies from a minimum of $0.01 to maximum of $999, in which at least half of the products has a minimum open bid amount of $4.50.

To predict whether the auction bid for a product is competitive or not, we have built a classification model. This model was developed by using decision tree classifiers. Since few variables are categorical, they have been converted to numerical values by using Label Encoding techniques and each value of these categories has been converted to a number.

	Category	Cat_z	count
0	Antique/Art/Craft	0	177
1	Automotive	1	178
2	Books	2	54
3	Business/Industrial	3	18
4	Clothing/Accessories	4	119
5	Coins/Stamps	5	37
6	Collectibles	6	239
7	Computer	7	36
8	Electronics	8	55
9	EverythingElse	9	17
10	Health/Beauty	10	64
11	Home/Garden	11	102
12	Jewelry	12	82
13	Music/Movie/Game	13	403
14	Photography	14	13
15	Pottery/Glass	15	20
16	SportingGoods	16	124
17	Toys/Hobbies	17	234

From the picture above, it's shown that each category has a number mapped under cat_z and number of their occurrences in the data.

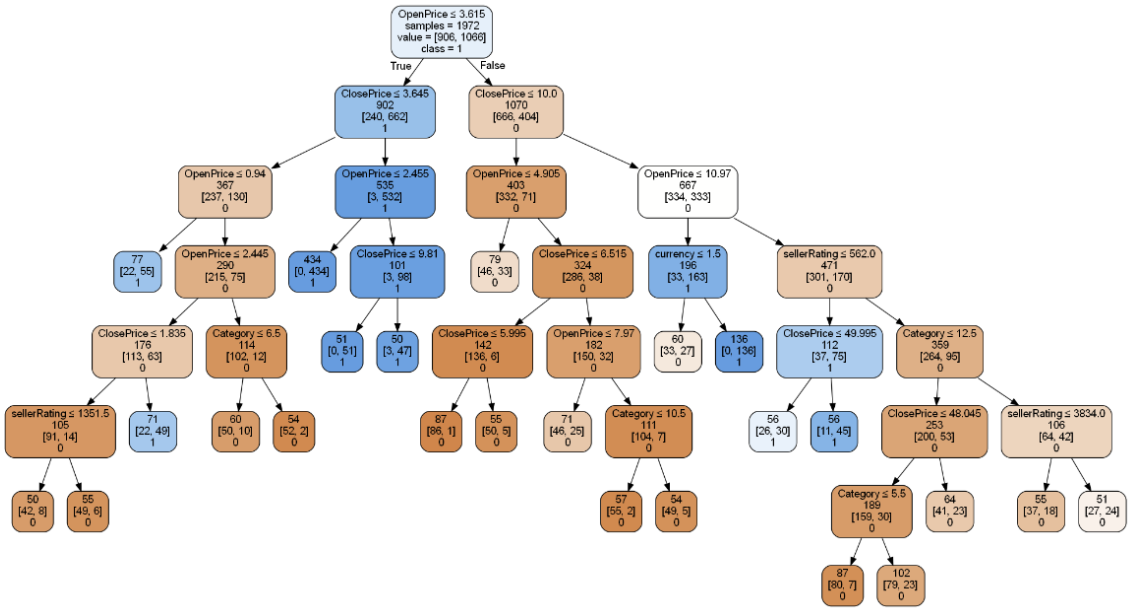
All the variables has been classified as predictors and outcome. These predictors have all the variables except the variable "Competitive" which is a target variable. Then this dataset has been divided into training data (60%) and validation data(40%)

A decision tree has been developed using the training data, with a max_depth of 7, which gives at least 7 levels of nodes and min_samples_leaf=50 which ensures that each node has at least 50 values.

So, upon building this model there are 7 features in this model which helps in predicting if the bid for a particular product was competitive or not. Then the model has been fitted using the training data. Multiple Rules can be made using the decision tree which was developed.

This classification tree has an accuracy of 84.6%

Out[77]:



By looking into the above classification tree, we can make a rule that if the open price of the product is less than $3.615 and if it was closed by greater than $3.645, and if the open price of the product is less than $2.45 then auction/bidding for that product is competitive.

Like the above rule, multiple rules can be made from using the classification tree which was developed.

But the above classification tree is not very useful for making a prediction for new product. As this tree consists of variables such as closeprice, endday which are not available for a new product while making a prediction and also this tree has variable such as currency which is not significant as the bidding is happening in only USD and openprice, closeprice is also in USD.

So, a new model has been developed by considering the variables which are available at the time of auction and helps in predicting the outcome for the target variable.

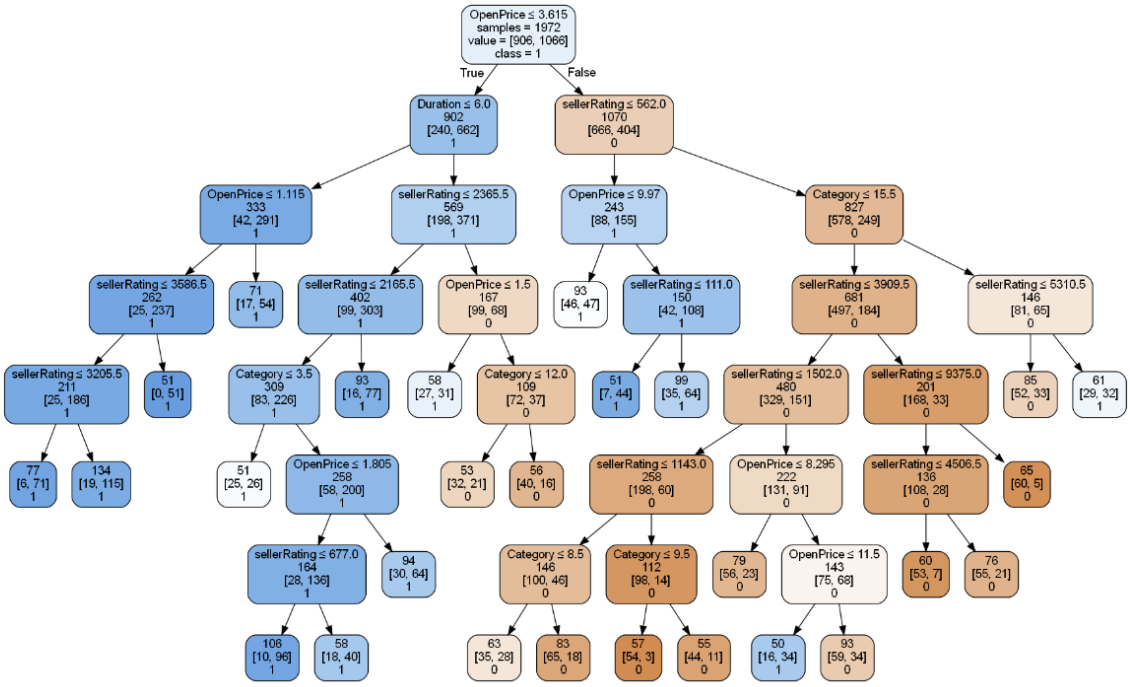
The variables which are included in this revised model are ‘Category','sellerRating', 'Duration', 'OpenPrice’ And the target variable remains the same (‘Competitive)

A new model has been developed with only the variables which are available at the time of auction and a revised decision tree was made for predicting a new auction bid. In this decision tree the same criteria as earlier tree was given i.e., max_depth =7 and min_samples_split=50, min_samples_leaf=50
By using these criteria, we have ensured that there at atmost of 7 nodes in the tree and each node has at least 50 nodes while making a split and at least 50 nodes in every node in the classification tree.

Accuracy of this revised model is 73.6%

If a seller has to make a decision on auction bid whether if the bid is competitive or not, they can consider looking into the decision tree and make a decision.

Out[84]:



To make a best bid in the auction, the seller can consider the following scenarios which can make them money for a product which is highly competitive.

To make the auction more competitive, set the product open price for less than $1.15 and set the maximum duration of 5 days for the product in auction. Make sure the sellerrating is less than 3587 then is a high chance that the product can be bid on by more than one buyer and hence seller can get more profits from this product.

Also, as bidding takes place in USD, it is better to use USD while pricing a product in the auction.