AllLife Bank

- PowerPoint Presentation -

CONTENTS

BUSINESS PROBLEM OVERVIEW

AllLife Bank is a US bank that has a growing customer base.

The management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors).

1.Which variables are most significant ?
2.Which segment of customers should be targeted more?

- Most of our customers are liability customers (depositors) with varying sizes of deposits.

- The number of customers who are also borrowers (asset customers) is quite small.

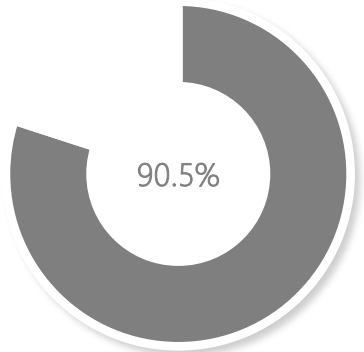
- The bank is interested in expanding this base (asset customers) rapidly to bring in more loan business and in the process, earn more through the interest on loans.

- A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio.

DATA OVERVIEW

Data

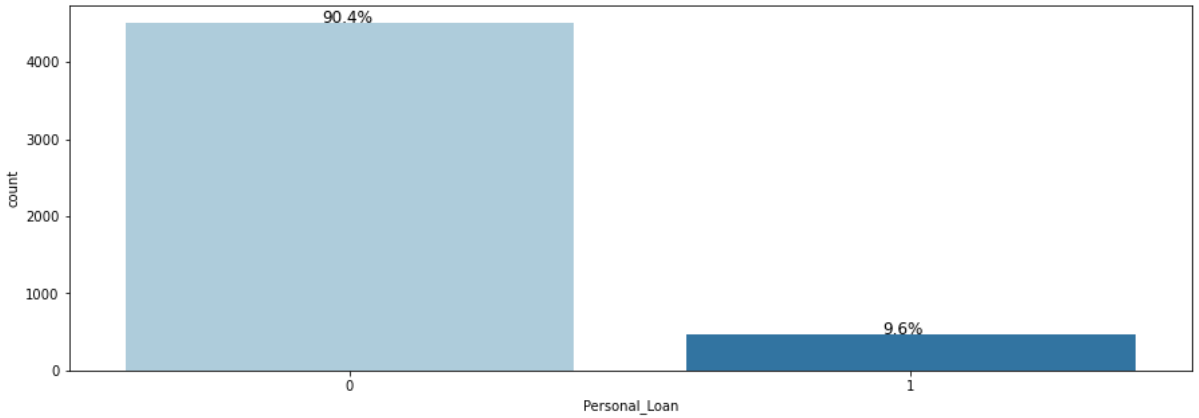
- The data contains information about 5000 customers and their 14 (rows) information as:
  - Id and Zip code;
  - Family (1, 2, 3, 4) and Education (1 = Undergrad, 2 = Graduated, 3 = Advanced/Professional) as Ordinal Categorical variables.
  - Binary Categorical variables such as Securities Account, CD Account, Online, Credit Card, Personal Loan (Our Target Variable).
  - Income, Age, Experience, CC_Avg, Mortgage.
- Zip Code we will treat it, classifying your customers in Unknown (34 Zip Code outside of EUA), North or South of California.
- Experience has negative values, which we will treat it before proceeding with the analysis.
- Some information are right skewed, and with some outliers, after analyzing it, we realize that is a real data, except for customers with less than 6y experience with an income greater than 185 thousand per year, so we will drop this 17 customers (rows).



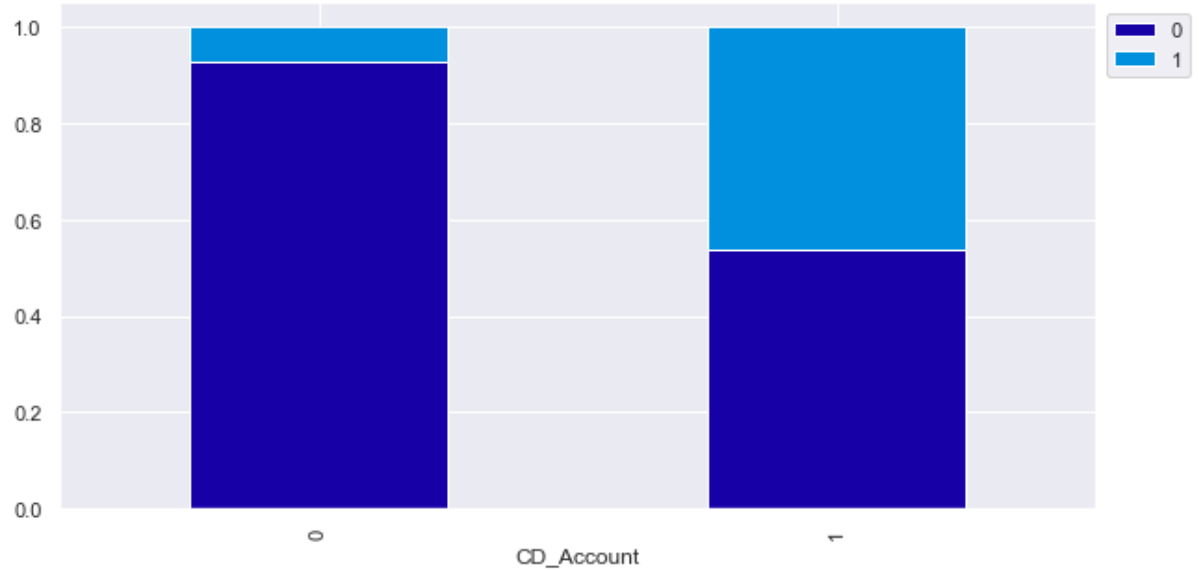
Personal Loan

We have an imbalanced data set, with 90.5% of customers as Won't by a Personal Loan and only 9.5% that Will by a Personal Loan.

EXPLORATORY DATA ANALYSIS

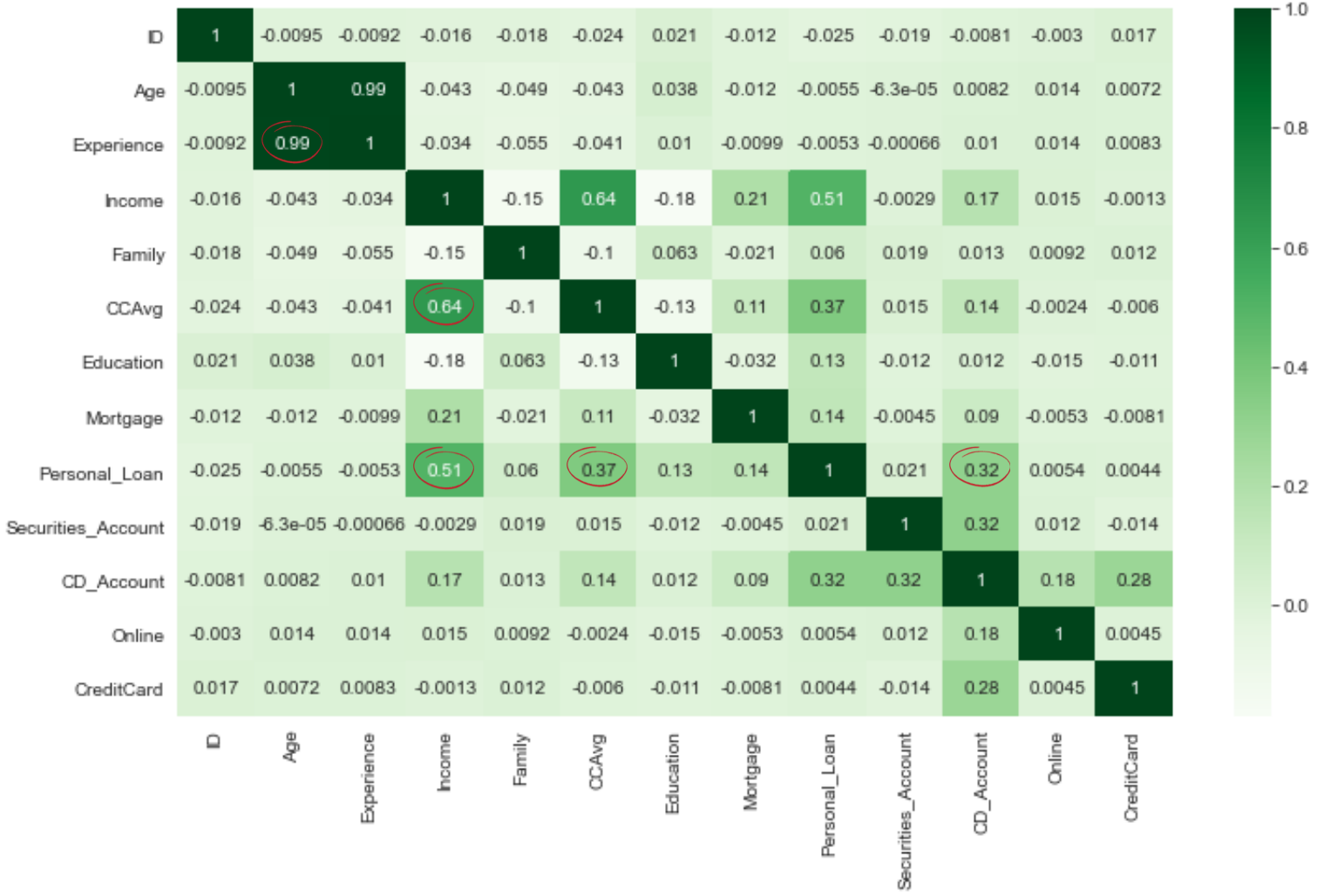


It's a imbalanced data where only 9.6% of customers said yes for a Personal Loan.



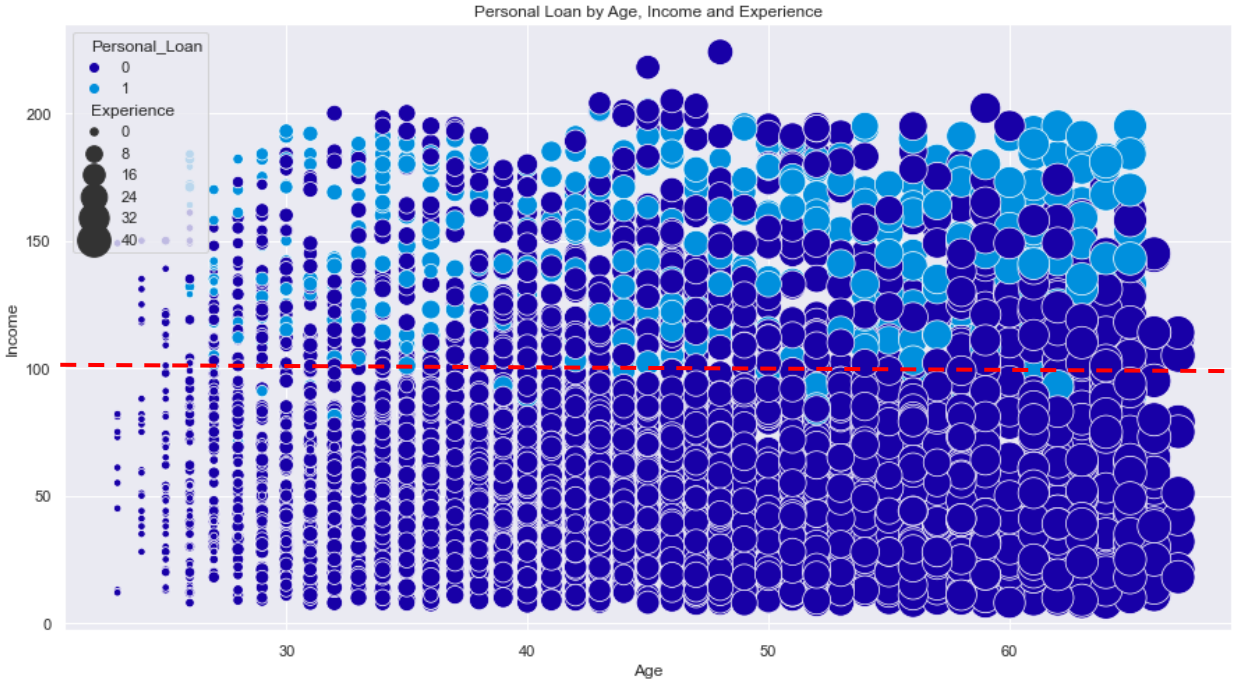
- Only 6% of customers have a CD_Account
- Customers with CD_Account is more likehood to accept a Personal Loan (46,17%)
- A good strategy is try to convert more clients to become a CD_Account.

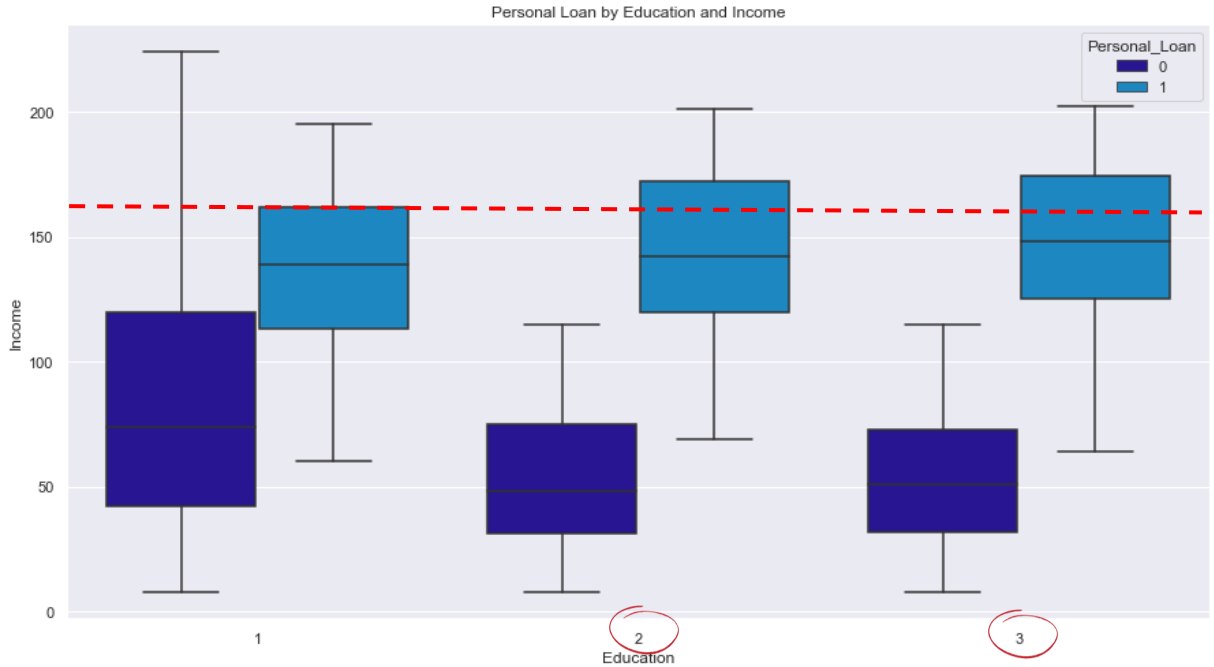
EXPLORATORY DATA ANALYSIS



- Experience shows the highest correlation with Age (0.99) which was expected.
- CCAvg has a 0.64 correlation with Income.
- Persolnal_Loan has 0.51 correlation with Income, 0.37 with CCAvg and 0.32 with CD_Account
- There is no relationship with Id and Personal Loan

EXPLORATORY DATA ANALYSIS

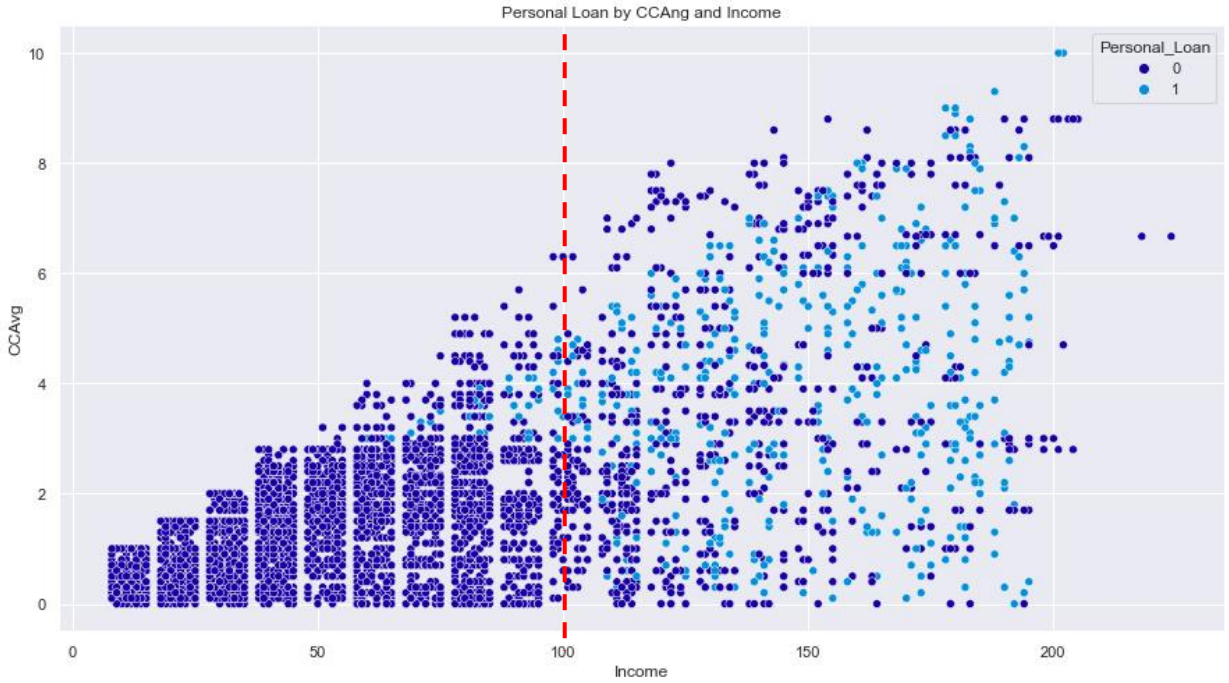




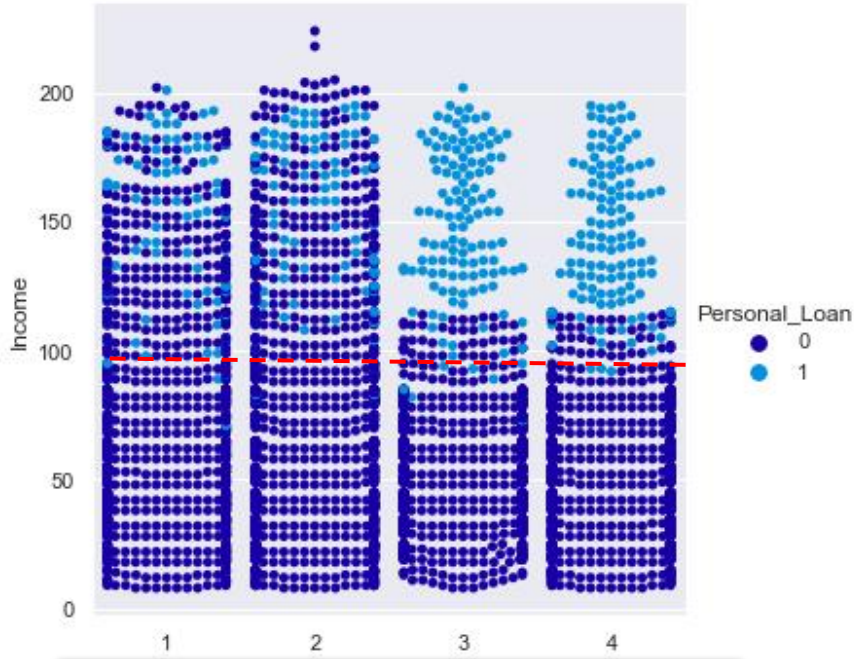
- Customers with highest income (greater than $100 K/year) tend to accept Personal Loan.
- Age and Experience seems to have a strong positive correlation.
- Age and Income shows some correlation.

- Customers with high Education is more likehood to accept a Personal Loan.
- 13.54% and 12.84% of customers with Education level 3 and 2 respectively accepted a Personal Loan.
- Customer with high Education has Income grater than 160k.

EXPLORATORY DATA ANALYSIS



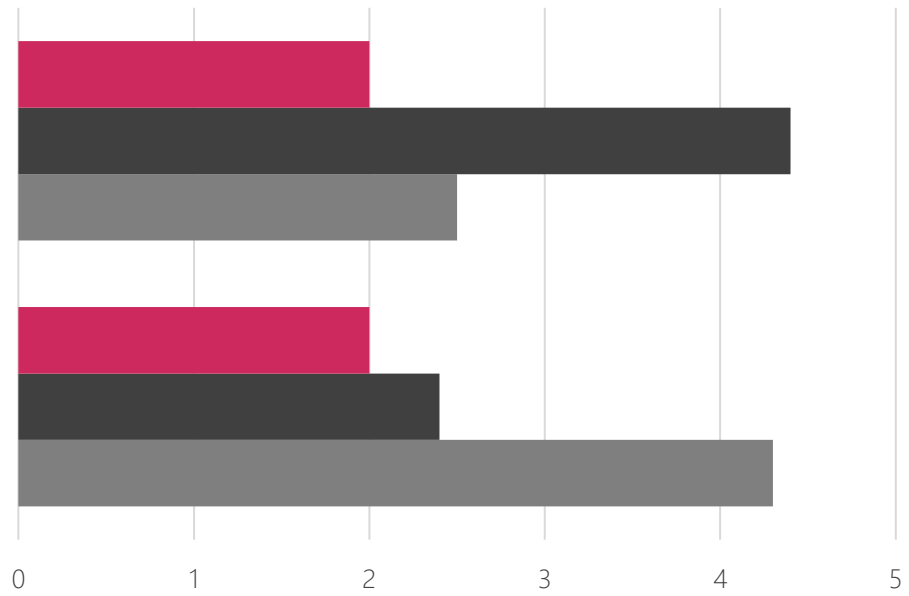
- Customers with high Income usually has high CCAvg, what is expect to happen.
- Higher Income, more likehood to get a Personal Loan.
- There is a wide range on CCAvg and Income, but doesn't seems to be outliers.



- Customers with high Family size and Income greater than 100 thousand is more likehood to accept a Personal Loan.
- 14.93% and 12.22% of customers with Family size 3 and 4 respectively accepted a Personal Loan.

ALLLIFE Bank

MODEL PERFORMANCE SUMMARY



We analyzed the "Personal Loan accepting" using different techniques: Logistic Regression and Decision Tree Classifier to build a predictive model for the same data set.

The built model can be used to predict if a customer is going to accept or not a Personal Loan and to create a Customer segment considering the significance of independent variables

MODEL PERFORMANCE SUMMARY

Logistic Regression

	Model	Train_Accuracy	Test_Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
0	Logistic Regression Model - Statsmodels	0.956709	0.945819	0.660606	0.602740	0.848249	0.792793
1	Logistic Regression - Optimal threshold = 0.08	0.879014	0.890970	0.906061	0.897260	0.433333	0.469534
2	Logistic Regression - Optimal threshold = 0.33	0.946961	0.945151	0.721212	0.705479	0.719033	0.725352

Conclusion

- Our Best Logistic Regression predictive model to find the segment of customers who will buy a personal loan with a Recall of 0.7054 on the test set and formulate devise campaigns accordingly.
- Threshold 0.08 is to low, which imply in high False Positives, we don't want to target wrong customers, we need a balanced Recall and Precision, even Recall is more important for us.
- Coefficient of some levels of Income, Family size, CCAvg (Average spending on credit cards per month), Education level and CD_Accouting (Customers with CD account with the bank) are positive an increase in these will lead to increase in chances of a person accepting a Personal Loan.
- Coefficient of Onile, CreditCard, Securities_Account are negative increase in these will lead to decrease in chances of a person accepting a Personal Loan.

MODEL PERFORMANCE SUMMARY

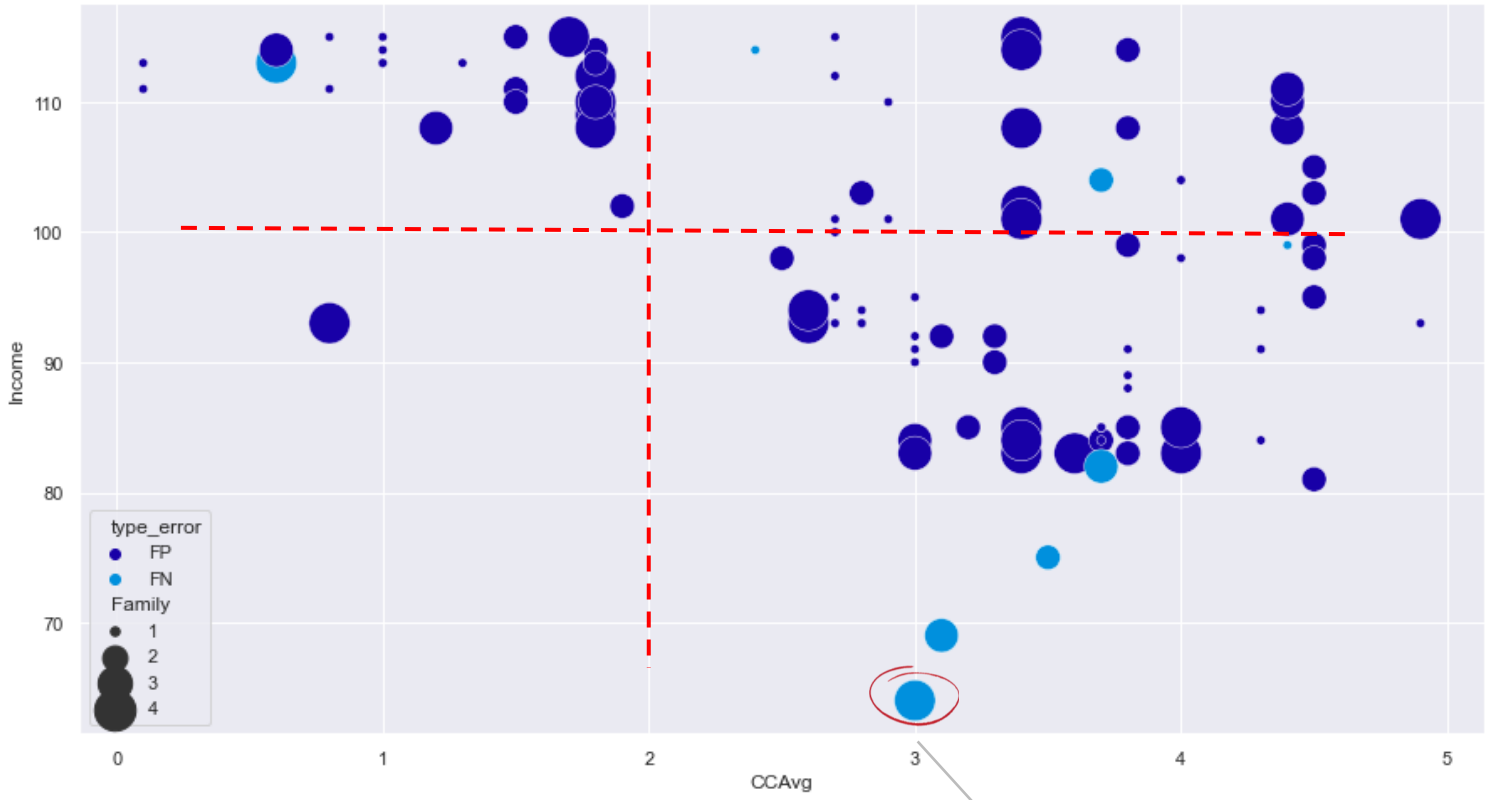
Decision Tree

	Model	Train_Recall	Test_Recall	Train_Precision	Test_Precision
0	Initial decision tree model	1.0000	0.8836	1.0000	0.9021
1	Decision treee with hyperparameter tuning	0.9879	0.9589	0.3804	0.4082
2	Decision tree with post-pruning	0.9939	0.9795	0.5899	0.6204
3	Decision tree with post-pruning - alfa 0.00135	0.9970	0.9521	0.8209	0.7989

Conclusion

- Our Decision Tree predictive model that can be used by AllLife bank to find the segment of customers who will buy a personal loan with a Recall of 0.9521 | alpha 0.00135 on the test set and formulate devise campaigns accordingly.
- Post-pruning with best model, has low precision which imply in high False Positives, we don't wanna target wrong customers, we need a balanced Recall and Precision, even Recall is more important for us.
- Feature Importances are Income, Family size, CCAvg (Average spending on credit cards per month), Education level and Age these are the Customer Segmentation that will lead to increase in chances of a person accepting a Personal Loan.

INCORRECTLY PREDICTED DATA



FP	107
FN	8

Total 115 Incorrectly Predicted Data

We can see that the error occurs due customer profile.

- Some customers even having a accepting Personal Loan profile (High Income, CCAvg, and family size) it does not accept.
- On the other hand, customers with no accepting Personal Loan profile, will say yes, Those ones we need to target then as well.
- FP customers, have a high income (>80k), with high Education (>=2) and High CCAvg (mean >=3k), but did not accept Loan.
- FN customers, have a High CCAvg (mean >=3.05k), but a min income of 64k and max 114k, with Education mean around 2 and family size around 2, our model fail in detect it as a accepting Loan customer because they significant variables is opposite of a regular accepting Loan customer.

CONCLUSION

- We visualized different trees and their confusion matrix to get a better understanding of the model. Easy interpretation is one of the key benefits of Decision Trees, followed by less data pre-processing (outliers, missing data, features engineering…). In the other hand Logistic Regression request all this data pre processing and it is difficult to correctly interpret the results and outliers affects the model. Logistic regression looks at the simultaneous effects of all the predictors, so can perform much better with a small sample size.

- We verified the fact that how much less data preparation is needed for Decision Trees and such a simple model gave good results even with outliers and imbalanced classes which shows the robustness of Decision Trees.

- The models predicted in logistic regression was 0.8973 on test and the decision tree was 0.9521 on test. The data sets did better with Decision Tree.

- Income, Family, CCAvg, Education, and Age are the most important variable in predicting the customers that will accept the Personal Loan

- We used post pruning to reduce overfitting and choose the alpha 0.00135 to get a best Recall considering a good Precision as well. Meaning that, Recall is the most important metric, but we also want to have a good precision to make sure we are targeting the correct customers.

BUSINESS INSIGHTS AND RECOMMENDATIONS

- According to the Decision Tree model, pos tunning bestmodel2:
- a) If a customer `Income` is greater than 100k, there's a very high chance the customer will accept the Personal Loan.
- b) If a customer has high Income and `Family` size greater or equal to 3, there is a high chance of this customer accept a Loan
- It is observed that the more the customers spend on CCAvg more is the likelihood of them to accept the Loan, also the bank can enhance the customer experience with the bank products as Securities Account, CD Account. Bank needs to do a strategi to convert more customers with CD Account, and CCAvg usually they will by a Personal Loan.
- After all campaign the model needs to be review (new information) and a new segmentation of customers should come out.
- Customer Segment is Income greater than 100k, Family size bigger than 2, High CCAvg and Higher Education.


THANK
YOU
- Amanda Mendonca -