PREDICTING CREDIT CARD BALANCE

Suzana Iacob

Example of prior analytics work

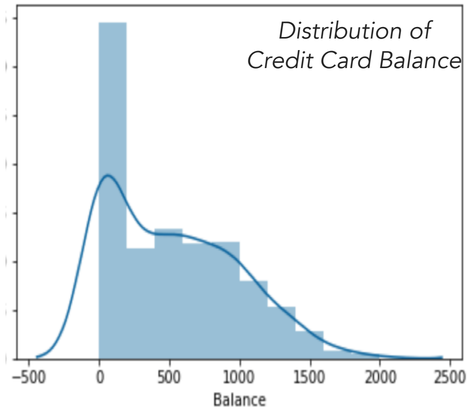
This presentation was prepared as a summary of an independent project I completed. Full project available at kaggle.com/suzanaiacob/predicting-credit-card-balance-using-regression
It was presented as part of the "10 Minutes Learning Series", an initiative I am involved in.

STATISTICAL ANALYSIS OF CREDIT CARD DATA

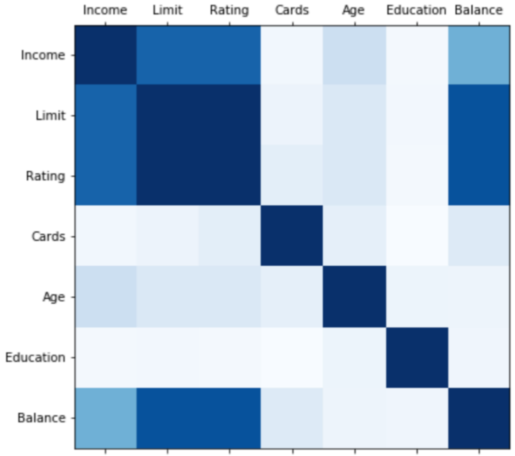
Purpose: To comprehend which factors influence the Credit Card Balance of a cardholder and to predict the average Balance of a given individual.

Analysis conducted on dataset of 400 observations and 10 variables, 6 of which numerical and 4 categorical

- Exploratory Data Analysis revealed a significant portion of sample as being **Zero Balance Cards**, leading to fitting an additional model for active-only cardholders
- Non-active cardholders could maintain a zero-balance to **decrease their credit utilization** and boost their credit rating, assuming they also own a positive balance credit card elsewhere
- A strong relationship was observed between Balance, **Credit Limit**, **Credit Rating**, and **Income**
- Initial analysis did not identify an association between Balance and Gender, Ethnicity, Education or number of Cards. Although these variables **do not appear significant when observed in isolation**, their interaction with one another might make them valuable
- **Limit and Rating** are highly correlated, introducing multicollinearity. Rating as an antecedent of Limit is more meaningful because it also drives credit Limit levels.



Distribution of Credit Card Balance



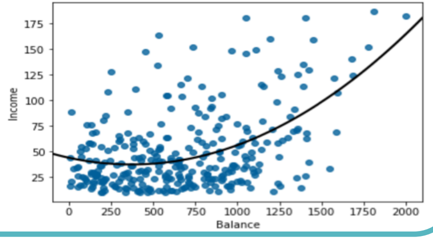
Correlation matrix among all numerical predictors and Balance

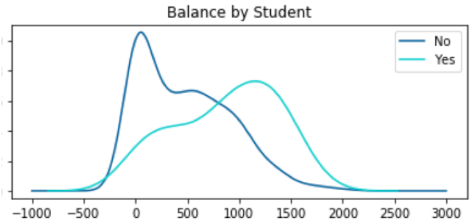
PREDICTORS FOR CREDIT CARD BALANCE

A **non-linear relationship** was observed between Income and Balance.

At lower levels of Income, increases in personal Income cause a decrease in credit card Balance, which can be interpreted as individuals **requiring less financing** as they make use of personal finances instead of credit debt.

At high levels of income, Balance increases; for those individuals loans are in higher demand, potentially due to **increased investment activities** and a greater risk tolerance.





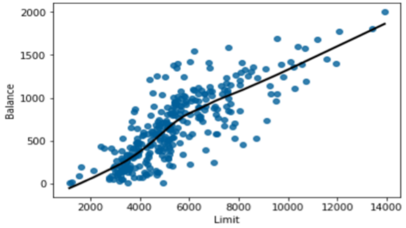
Students display on average **higher** credit card Balances. We infer that **Students have higher need for financing** due to student loans and lower income, hence their financial flexibility may be lower.

Further analysis indicated that increases in **Income level** cause increases in the Balance of **non-Students**, however, in the case of Students, changes in Income do not impact their average Balance.

Credit Rating and Credit Limit appear to be the strongest predictors for credit card Balance, each explaining 74% of the variance in balance.

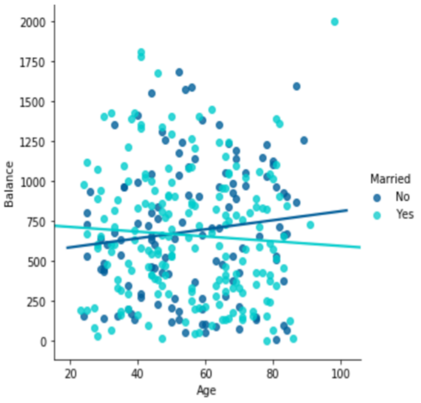
This could suggest that individuals with **high Rating are more willing to incur credit debt** as they are confident that they will be able to pay off the balance.

Both Limit and Rating are **complex measures**, subsuming a range of other factors, among which several already present in the model, such as Income.

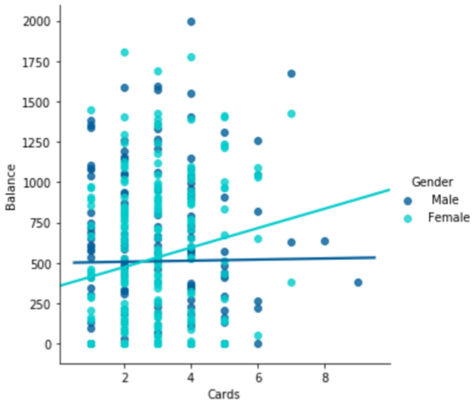


Project built on Kaggle using Python 3. Visualizations created with Matplotlib and Seaborn

INTERACTION BETWEEN VARIABLES AND FINDING THE BEST MODELS



While neither Married nor Age are significant in isolation, the interaction term is. This implies that individuals with higher values for **Age** who are also **Married** have lower credit card Balances pointing to **higher financial prudence or risk aversion**.



Females who own **more Cards** have on average a higher Balance.

Gender in isolation has a negative impact on Balance, suggesting that females, in general, have less credit card debt, except when the individual also owns multiple Cards.

Best Models

1. **Entire dataset:** Balance ~ Income + Income**2 + Age + Student + Limit + Cards + Income*Rating
2. **Active cardholders dataset:** Balance ~ Limit + Rating + Income + Age + Student + Cards
3. **Active as Binary Outcome:** Active ~ Income

For the **entire dataset (1)**, the best model predicted **96% of the variance**, while the model fit on the **active-only population (2)** predicted **99%.** The difference suggests that there are other factors influencing non-active cardholders which are not present in our data, or their spending behaviour is reflected on other lending platforms.

Using Logistic Regression, the **Active outcome (3)** was best predicted by **Income**, with high earners having a greater probability of being active. This could be explained by low earners maintaining a zero-balance card in order to boost their credit worthiness.

Regression models built with StatsModels

THANK YOU