**Summary:**

**This is a banking data set. The target variable is First Payment Default. Data cleaning, feature engineering, model selection, threshold value, evaluation everything is visible in the Jupyter notebook file. This summary consists of quick overview about the project outcomes.**

**1.The customers are more from CA than TX.**

**CA-> 1097**

**TX-> 357**

Chart, bar chart, box and whisker chart

Description automatically generated

**2. Payment Default ratio is high in Texas which is 35 as compared to CA which is 23 percent approx**

Text

Description automatically generated with medium confidence

**3. Customers are more from rent than own**Chart, bar chart

Description automatically generatedn.

**4. Payment Default ratio is high from own which is 29 percent as compared to rent which is 26 percent. So this means 29 own customers out of 100 are defaulting their first payment.**

**First Payment Default Rent or Own**

**False R 1038**

**O 137**

**True R 368**

**O 57**

**Name: Rent or Own, dtype: int64**

**5. Interestingly the first default payment is neither strongly positively correlated nor negatively correlated with dependent variables. i.e there is no strong one factor that is impacting the default payment.**

Chart, treemap chart

Description automatically generated

**6. Monthly Net Income has high number of outliers among all the features.**

A picture containing graphical user interface

Description automatically generated

**Customers whose income are different from rest i.e either too high or too low. Here are their values:**

**Most of the income fall under 9870.0**

**2800.0,**

**10500.0,**

**11400.0,**

**10000.0,**

**15000.0,**

**20000.0,**

**25000.0,**

**30000.0,**

**10800.0,**

**14000.0,**

**10032.0,**

**19000.0,**

**16188.0,**

**11200.0,**

**14400.0,**

**184000.0,**

**11460.0,**

**14150.0,**

**13000.0,**

**13008.0,**

**18000.0,**

**11608.0,**

**10200.0,**

**10334.0,**

**12000.0,**

**21732.0,**

**10600.0,**

**12400.0,**

**12018.0,**

**11000.0,**

**9980.0]**

**According to the observation, these outliers doesn’t have any specific outcome on the first payment default.**

**(please refer to jupyter notebook for more info).**

**7. ROC curve :**

**Looking at this curve the domain expert can decide whether to increase or decrease the True positive rate and False positive rate.**

Chart, line chart

Description automatically generated

**8. Highly imbalanced dataset. Most of the customers were not first payment default.**

**Chart, bar chart

Description automatically generated**

**The machine learning model was up sampled to balance it out.**

**The final test data is predicted at its best. The precision and recall metric was considered. Random forest came out to be having the best values of precision and recall. The threshold was calculated by taking the mean of 4 different algorithms and the best threshold corresponding to the maximum accuracy was considered. The values are ranging between 0.53 to 0.58**

**Table

Description automatically generated**