**Capstone 2 - Milestone Report 1**

**Default on Home Loan**

**Problem**

No one wins when a loan is defaulted on. The borrower sees a significant drop on their credit score, faces the potential of losing their home, and an immeasurable amount of stress. Though the lender may recoup some of their money by foreclosing on and selling the home, the lender typically has to sell the home for less than it is worth (they lose money every day and must encourage someone to buy the property as soon as possible) and also loses out on the money they would have earned in interest. Lenders screen loan applicants vigorously, but many are still able to secure a loan and soon fall behind on payments. To save both the borrower and lenders from the horrors of a foreclosure it would be beneficial if lenders could tell which clients may have trouble paying their loans early on. To this effect I will construct a model to predict if an applicant will miss one or more of their early loan payments.

**Clients**

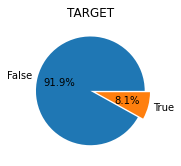
A model that predicts if an applicant will miss one or more of their early loan payments would be very beneficial for home loan lenders. Borrowers who miss early payments made it through the initial screening and are either good clients who may need some additional attention/assistance or are fraudulent. The proposed model will allow lenders to flag accounts for targeted efforts such as initiatives to make sure the client is thoroughly informed, programs that monitor and assist borrowers sooner, or additional screening to determine if an application is fraudulent. The model may also be beneficial to borrowers as it can be used in conjunction with other informative applications and services (i.e. “Can you afford this mortgage” calculators) by the more proactive and risk averse borrowers.

**The Data**

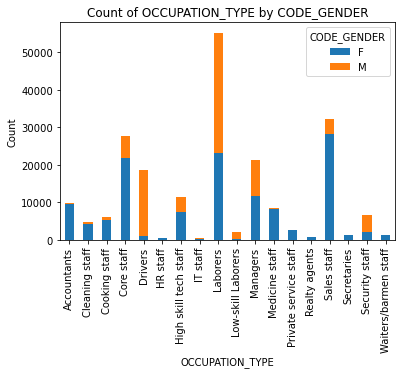
The data was download from Kaggle in CSV format (<https://www.kaggle.com/mishra5001/credit-card?select=columns_description.csv>) then imported into Python and analyzed. There were 307,511 unique records consisting of loan recipients. The data set contained 120 features and the target variable “TARGET” which indicated if a recipient missed one or more of the first few installment payments on their loan (binary true/false.) The features indicate what type of loan the recipient received and a wide range of demographic information including the recipient’s age, details regarding where they currently live, etc. (at the time of application.) Information regarding the recipient’s current dwelling have been normalized.

Features with more than 20% of their records missing were dropped as replacing these missing records with some value may lead to some misleading results in the model. Records that were considered “missing” if there was no data, “XAP”, “XNA”, or 365243 was present 365243 was present in fields measuring the days since an event, but this value is an outlier that does not correspond to any real meaning (i.e. attempting to interpret this value for the applicants age in days would mean that the applicant was approximately 1000 years old.) These values were replaced with “NA” for object variables, the arithmetic mean for integer variables, and the mean for continuous variables.

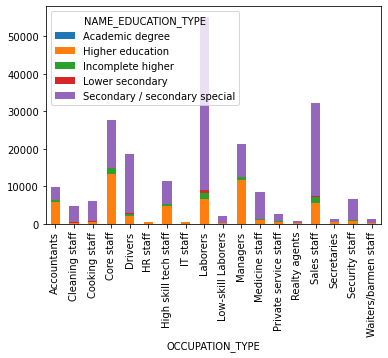
The target variable “TARGET” is highly imbalanced with only 8.1% of recipients having missed one or more of the first few payments. This may cause issues with training a model as many algorithms assume balance (i.e. the majority class may overshadow the minority class and perform poorly when predicting the minority class.) To combat this, I will also test algorithms that do not assume balance or can be tuned to compensate and also test applying techniques that reduce the effects of imbalance (i.e. oversampling.)



The female applicates outnumber the male applicants 2:1 with men being more prevalent in the driver and laborer occupation types.



Interestingly, the overwhelming majority of applicants have completed secondary education at most and are highly prevalent in most of the occupation types.



There are a number of high correlations between pairs of feature; however, many of these are in easy to interpret groups. For example, there is high correlation between the features and measurements regarding applicants’ current living arrangemt which may indicate that the applicants may be in the same geographical region and therefore homes have a common set of features or layout. There are no high correlations between the target variable and any of the features. The highest correlations with the target variable are negative correlations with the with the external data source variables (“EXT\_SOURCE\_1”, “EXT\_SOURCE\_2”, “EXT\_SOURCE\_3”) which also appear to have a similar positive correlation with one another.

