Home Mortgage Prediction

Team members:

Fariha Imam	fi2183
Jesse Woo	jw4202
Lewis Wu	zw2783
Sai Chintalapati	vhc2109
Sushant Prabhu	ssp2202

Background

Securing funds for **buying a home** is a pivotal financial endeavor, requiring **multiple loan applications**

We can **analyze home loan applications** to unlock insights in the home-lending market

Explore **effect of various factors** (e.g. applicant demographics) on **loan application approval**

Estimate the likelihood of home mortgage approval **using ML models**

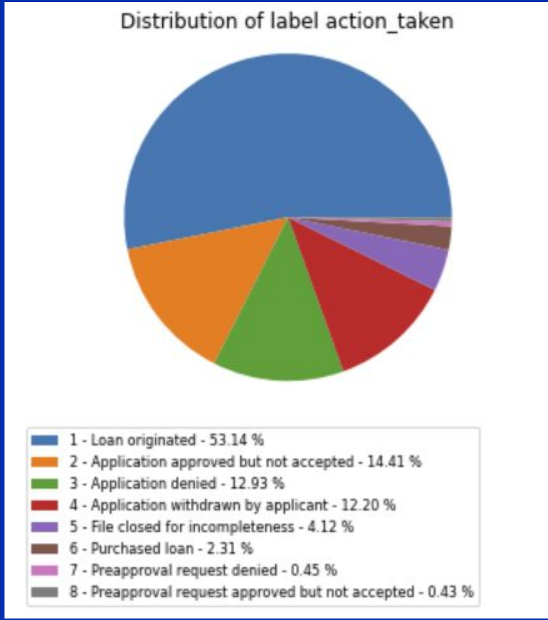
Streamline home buying process by offering predictive analytics to home loan applicants

National HMDA Dataset

- Comprehensive, publicly available data on the U.S. mortgage market
- Required by the HMDA (Home Mortgage Disclosure Act), and offered by the CFPB (Consumer Financial Protection Bureau)
- >5 million records per year, spanning 2007-2021
- Financial info on loan applications such as amount of credit granted, amortization rate, applicant demographics (i.e, race, age, sex, income), etc..

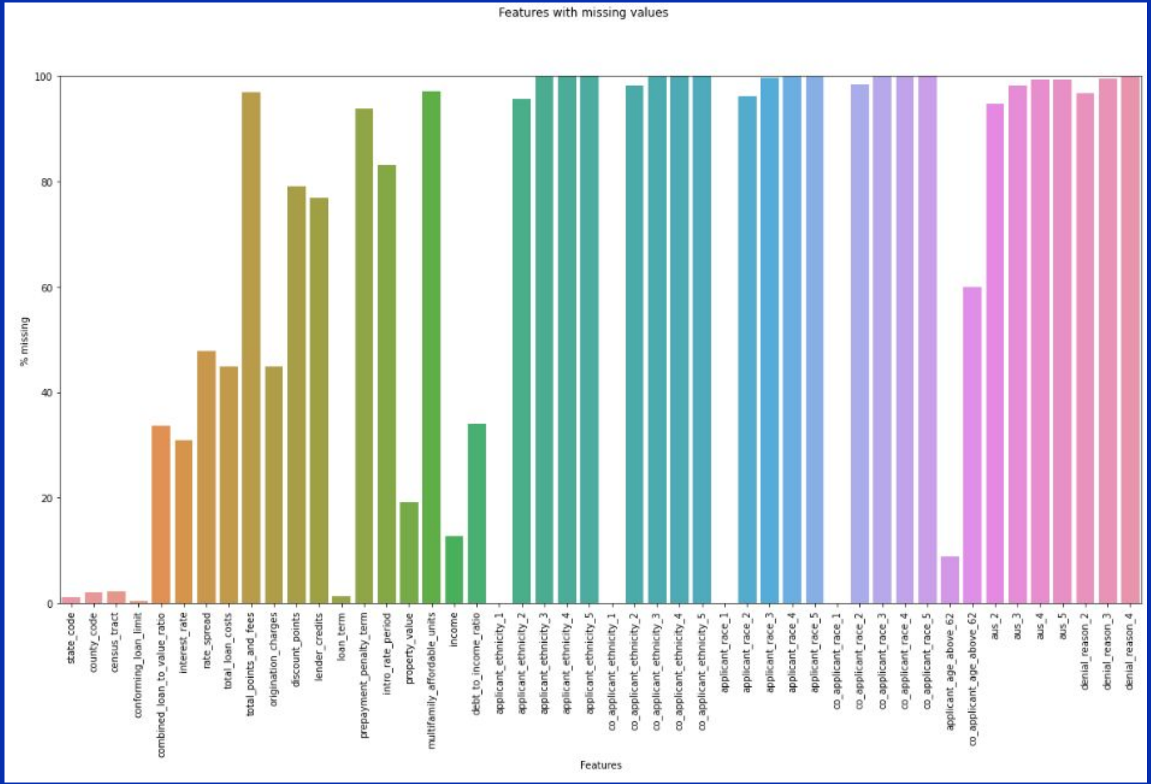
Initial Data Exploration

- The HMDA dataset contains 98 features
  - Categorical: 69
  - Numeric/Alphanumeric: 29
- Some numeric features are depicted with buckets (e.g. age category as 25-34)
- There are several fields that are aggregates of others
- The dataset contains one output class, describing the action taken on the loan application
  - The most common action taken is the origination of the loan, followed by application approved but not accepted, and application denied
  - It is very rare for a pre-approval request to be denied, and a pre-approval request approved but not accepted
- Given the very large number of features, a few important ones were chosen for EDA



Loan Application Outcome Classes

Initial Data Exploration



% of missing data for features with missing data

- There are 48 features with any number of missing values
- Many of the features have a very large percentage of missing values (over 90%)
- Missing values are often present because the features are optional
- Missing data took the form of null values, or some placeholder value (age = 8888)

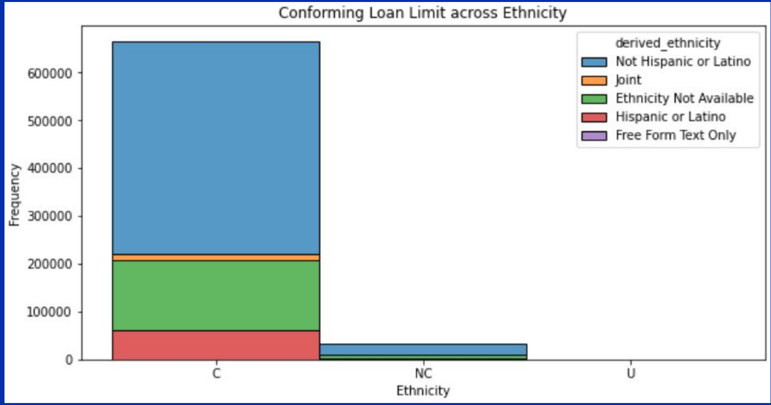
Cleaning and Sampling

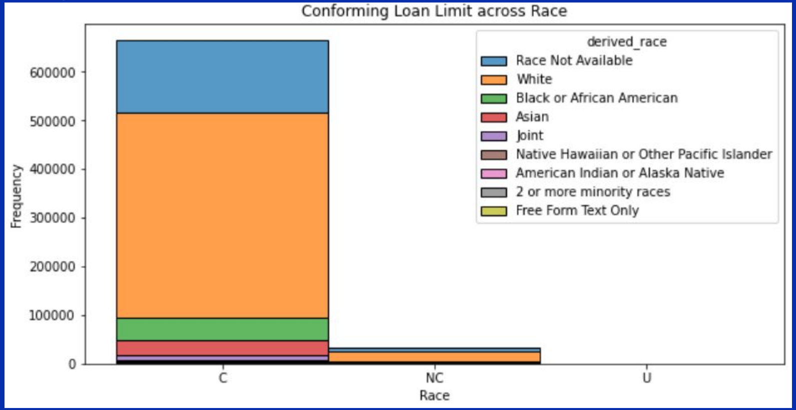
- The dataset initially contained 98 features
- 31 features with >50% missing data points were removed
- 4 features were taken off to prevent data leakage
  - Feature information gave insight into the reason an applicant was denied
- 33 additional highly correlated features (as determined by correlation matrix) were also dropped
- For important features where a small percentage (>10% but <50%) of data points were missing, the missing observations were dropped to preserve the feature
- For features with very few data points missing, the data was imputed using the mean value, where possible
- Sampling was required because the dataset is very large
  - Random sampling of observations was employed to increase generalizability

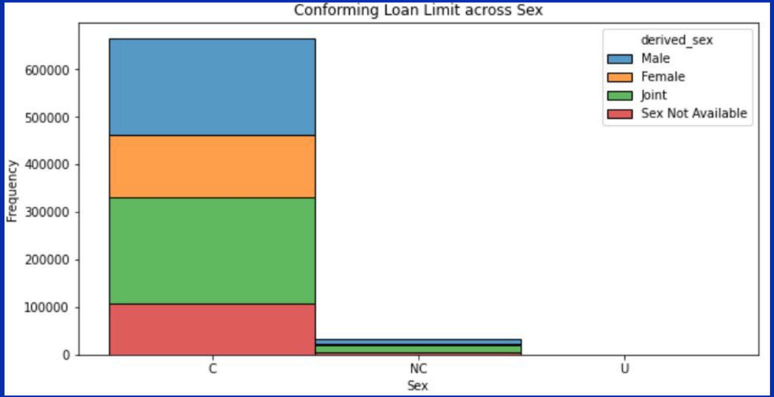
Takeaways - Resulting transformed data contained 501,586 observations and 30 features

Data Insights (EDA)

Loan Limits across Race, Ethnicity and Sex





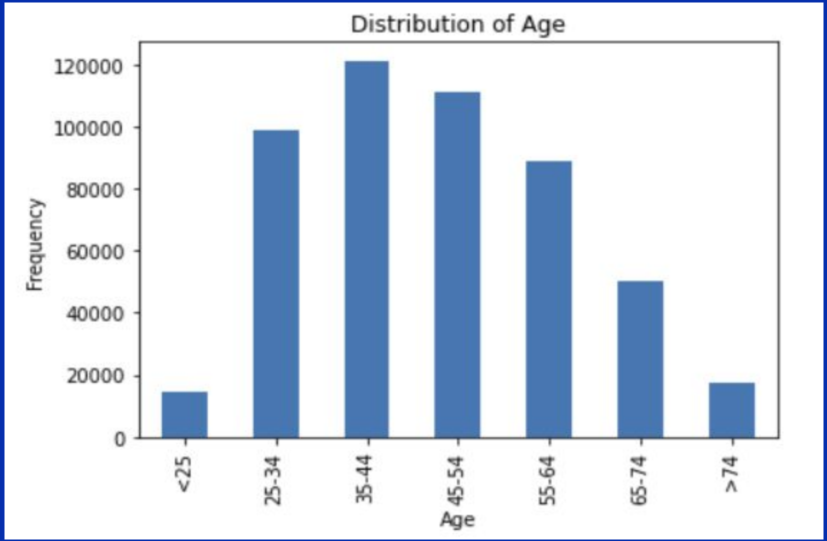


The conforming loan limit is the maximum amount of a loan governmental lenders are willing to guarantee. If the loan is greater than this amount, the loan is considered non-conforming.

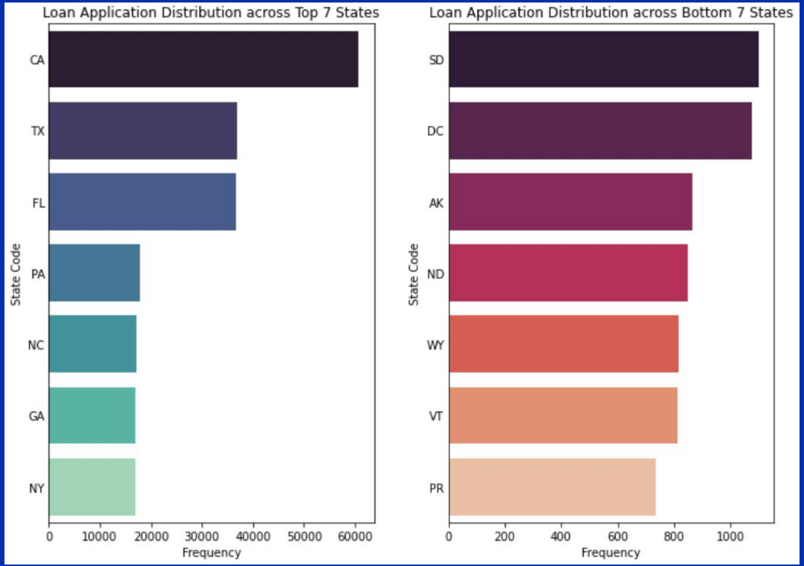
Takeaways -

- Majority of White people conformed the loan limit.
- Ethnicity being non Hispanic or Latino had highest share of Non-conforming loan limit category.

Data Insights (EDA)



Loan Borrower Age distribution



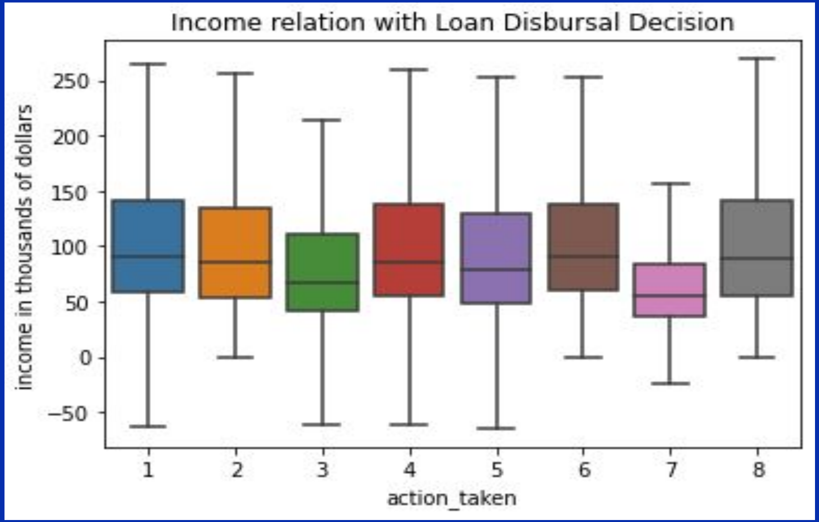
State Wise Applications (Top 7 & Bottom 7)

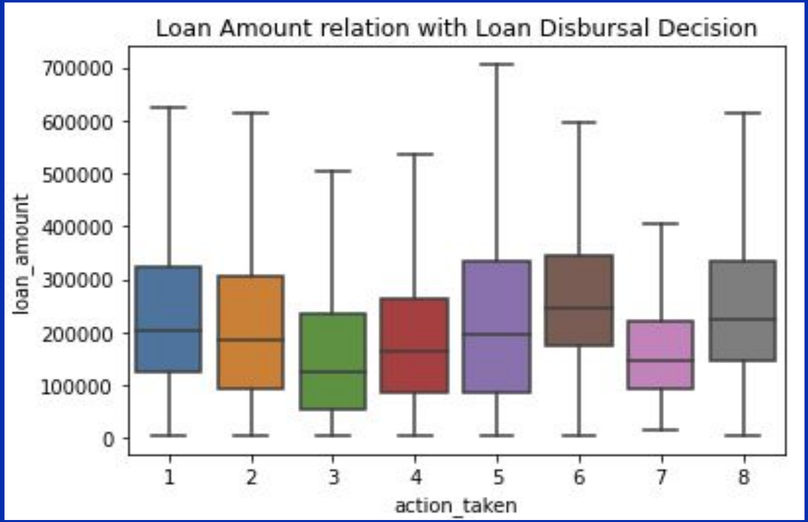
Takeaways -

- There are very few home buyers in the <25 age range, while the most common bucket of ages for home buyers is 35-44
- California had the highest number of loan applications (~60K) followed by Texas & Florida
- Wyoming, Vermont & Puerto Rico had the lowest, with Puerto Rico having merely 734

Data Insights (EDA)

Relation of Income & Loan Amount to Decision*





action_taken

1 - Loan originated (approved)
2 - Application approved but not accepted
3 - Application denied
4 - Application withdrawn by application
5 - file closed for incompleteness
6 - Purchased loan
7 - Preapproval request denied
8 - Preapproval request approved but not accepted

Takeaways -

- Loan originations skewed toward higher incomes and higher loan amounts
- Loan amounts <100K showed higher rejection rate
- Denials and purchase loans are both centered around lower incomes and loan amounts with a significant right skew

*Outliers removed in these visualizations

Machine Learning Techniques & Architecture

Machine Learning Models to Experiment -

- Logistic Regression
    - Serves as the baseline model; preliminary model assessment
- Decision Tree/Random Forest
    - Gives well-defined feature importance
- Support Vector Machine
    - Identifies clear margin of separation between classes
- Gradient Boosting/XGBoost/Adaboost
    - Provides high prediction accuracy
- Neural Networks (TBD to explore)

Planned Hyperparameter Tuning - (Grid search + K-Fold)

- LogReg:
    - Regularization type (L1, L2), C (Regularization parameter/Penalty strength)
- Tree-based methods:
    - max_features (Number of features to consider to decide the split), max_depth (Max depth of tree), n_estimators(Number of trees in the forest)
- Support Vector Machine:
    - Kernels type (linear, poly, rbf, sigmoid), C(Regularization parameter/Penalty strength)
- Boosting methods:
    - Loss (The loss function), Learning rate

Insights and Key Takeaways

Large data set with a high number of categorical and numerical features

Exploratory Data Analysis of target variable against application features

Many features dropped because of missing data or high correlation

Predicting loan outcomes is a multiclass classification problem with skewed and imbalanced targets. Will change to single class

Use logistic regression to establish a baseline

Use tree-based methods for classification and feature importance given number of categorical variables

SVM and NNs will also be considered, but model complexity is a consideration

Home Mortgage Prediction Modelling

Thank You!