**Midway Report**

**1. Question Formulation:**

Can I predict the median income of single moms for different regions in California based on socio-economic and demographic features?

- this changed from my original question because I realized there was a lot of similarity between the affordability ratio and other predictors, which made it a less interesting problem to solve.

**2. Data & EDA:**

* Has your data changed (more features, difference sources, etc) from your Nov 1 EDA? If so, provide summary statistics for any new data.

A table with numbers and a number on it

Description automatically generatedA table with numbers and letters

Description automatically generated

**3. Data Analysis and Model Building (Preliminary Stages):**

I have made a numeric transformer and categorical transformer for KNN and linear regression. The numeric transformer imputes missing values and fills with the mean, adds polynomial features, and scales the features. The categorical transformer also imputes missing values by using `most\_frequent`, then I OneHotEncode. Both of those are passed through my preprocessor. I also scaled my y’s because my target variable is income, which would give easier MSE values to interpret in the log form. For my decision trees I needed to make a new numeric preprocessor which only imputes missing values, replacing them with the mean.

* What insights or patterns have you observed during your initial data exploration?

I went through most of the steps without changing the format of income, so after getting very high and hard to interpret MSE scores, I decided to take the log of the income to fix that problem. I also changed my response variable because during my data exploration, I noticed `affordability\_ratio` was completely correlated with `CA\_RR\_Affordability`, and had very high correlations with other variables as well. This made the models I tried not very interesting or insightful, which is why I decided to switch to predicting income.

* If you've started feature engineering, which features have you created or modified, and why?

I created latitude and longitude features because I was hoping that those could be useful in some EDA by creating a map that can be used to get a good understanding of where the counties referenced are located. I also made certain features (region\_code, race\_eth\_code, geotypevalue, county\_fips) factors becauses they were originally numeric but since they represent different categories I changed their types.

* Which machine learning models or techniques are you currently considering or have already started implementing?

I have implemented simple KNN Regressor, Linear Regression, and Decision Tree models so far with some cross-validation, Ridge, and Grid Search used to try tuning my hyperparameters.

* Provide a specific list of elements you are incorporating in your project (from Item (D.) on [project description page](https://learningsuite.byu.edu/cid-23xW5FfSliq5/pages/id-c8dO))
  + Feature engineering for text data
    - one hot encoding certain features like county, race, and region
    - still thinking of another way to add a feature that would also make sense and improve my models
  + at least 4 supervised models: linear regression, KNN regressor, decision trees, XGBoost
  + deep learning model wit hyperparameter optimization
  + use SHAP to interpret feature importance
  + apply dimension reduction techniques
  + cluster analysis

**4. Challenges:**

* What has been the most challenging part of the project so far?

The most challenging part of the project so far has been deciding which kinds of models would make the most sense for my data, and the feature engineering process/EDA because I realized a lot of the features either don’t make sense or aren’t as useful as I thought they would be. I also feel like I need to include more non-tabular data for the feature engineering requirement or focus on a different topic to fulfill requirement (D) for the project.

* Have you had to pivot or adjust any aspects of your project based on unforeseen challenges or insights?

I had to change my target variable because I did not realize how highly correlated it was with some of the other features. I also might have to include a new topic from part (D) so that I can hit the grade requirement I want if I cannot find a way to do more non-tabular feature engineering.

* Are there any areas where you might need additional resources, guidance, or clarification?

I might need help with interpreting MSE values and how to make conclusions about my models, or how to work on some of the topics that I would need to study on my own.

**5. Collaboration and Feedback:**

* If you've sought feedback from peers, mentors, or others, what feedback have you received, and how have you incorporated it?
* If you haven't sought feedback, what is your plan to get feedback?

My plan is to work on the foundation of my project a little more and then once I have a better idea of what I want to do, I am planning on getting help on how to improve my analysis or add more complexity to it.

* How are you ensuring effective collaboration if the project is being done in a partnership setting?