Age Predictor

Data Preprocessing

- The data has been loaded from the csvs and converted to data frames. The data frames containing the job titles and the seniority for the respective job title has been merged on user ids and job titles. Then the **first job** for every user alongside its seniority has been filtered from this data frame to obtain **the start year of the first job** and furthermore create an important parameter: **number of years in the workforce.** Only the first job has been taken into consideration because the jobs after the first job don't make a difference to the number of years in the workforce.

- The same procedure has been implemented for the education data frame. Only the first major has been taken into consideration because the majors after the first major don't make a difference to the number of years since a user began his education. The first major for every user has been filtered out from the data. The parameter created is the **number of years from the start of the first major.** The type of major is bucketed into the four categories: **High School, Bachelors, Post Graduate** and **Other(the majors that cannot be bucketed into the other categories with surety).**

- **One Hot Encoding** has been used to include the buckets of the above categories as features. Bucketing the first major of every user into these buckets and one-hot encoding has provided better results.

- Close to 30% of the majors are recorded in a language other than English. These have been translated to English with the help of a function **translate_major**, written in the code. This ensures we do not lose out on important data and bucket these majors into their respective categories instead of 'others'. This piece of code takes longer to run,hence it has been commented out, an efficient method can be a future consideration.

- Lastly, the above data frames have been merged into a single data frame. The date columns have been converted to years and have been dropped later. Another column **'number of jobs'** the user has worked on has been created. The intuition behind this is that there could be a linear relationship between number of jobs and age. A column for the **number of years of the first major** has also been created. The intuition behind this is that if the category of the major is unknown, then the model can learn from the number of years of the first major.

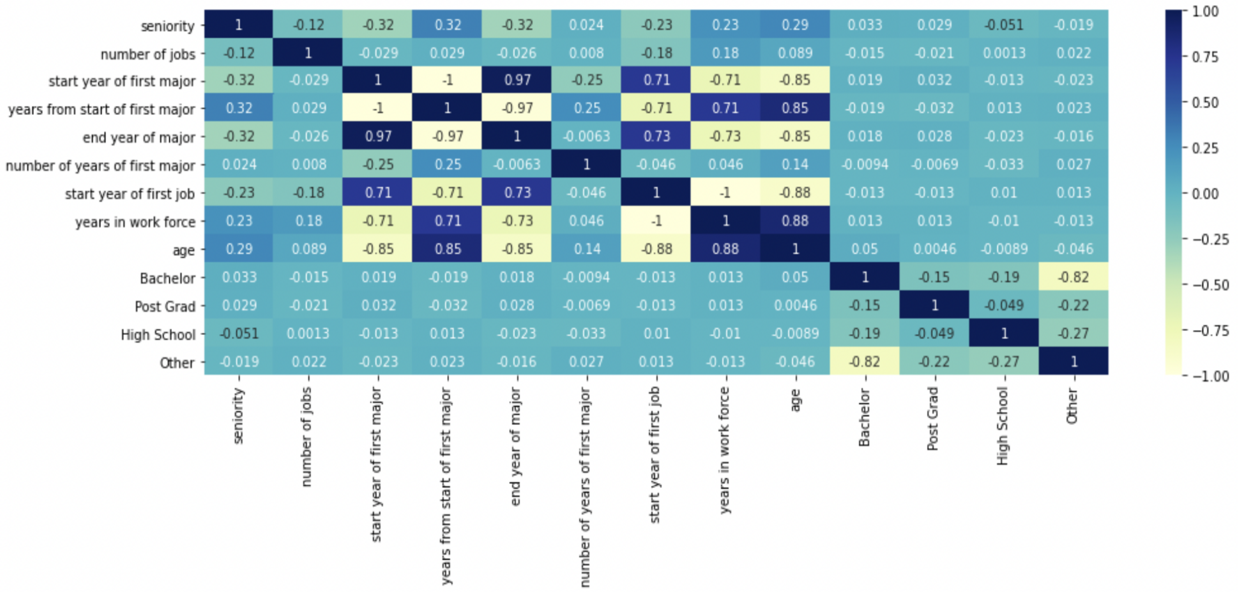
Establishing the Ground Truth

- The ground truth has been established primarily making the following assumptions:

    - The average age of a person when he starts his undergraduate education is **18.**
    - The average age of a person when starts his high school education is **16.**
    - The average age of a person entering the workforce is **22.**
- The above assumptions have been used in the following cases:
    - The first major start date has been used to determine the age of the user where the first major of the user can be bucketed as Bachelors/Undergraduate or High School. If the first major of the user is Masters/ PostGraduate/ a major that cannot be bucketed with any certainty the start date of the major is not used. This is because a person can start his Masters/ PostGraduate/ a major that cannot be bucketed with any certainty at any age between 20-30 or even more.
    - The first job start date has been used to determine the age of the user in the above case where the first major of the user is Masters/ PostGraduate/ a major that cannot be bucketed with any certainty.
- The training data is a **sample** of the population data. This sample should be **representative and random.** Hence for it to be representative, random **50%** rows are taken from each of these buckets: High School, Bachelors, Post Graduate and Other(the majors that cannot be bucketed into the other categories with surety) with their ground truths established accordingly.

Training Data EDA and Preprocessing

- The heat map created shows the relationship between different features of our training data. It is clear from the heat map that **age** has a strong relationship(correlation) with the **end year of the major, number of years since first major, number of years in the workforce and the years since the first job**. This is the relationship introduced by us as the ground truth and other data points can be observed to behave similarly.



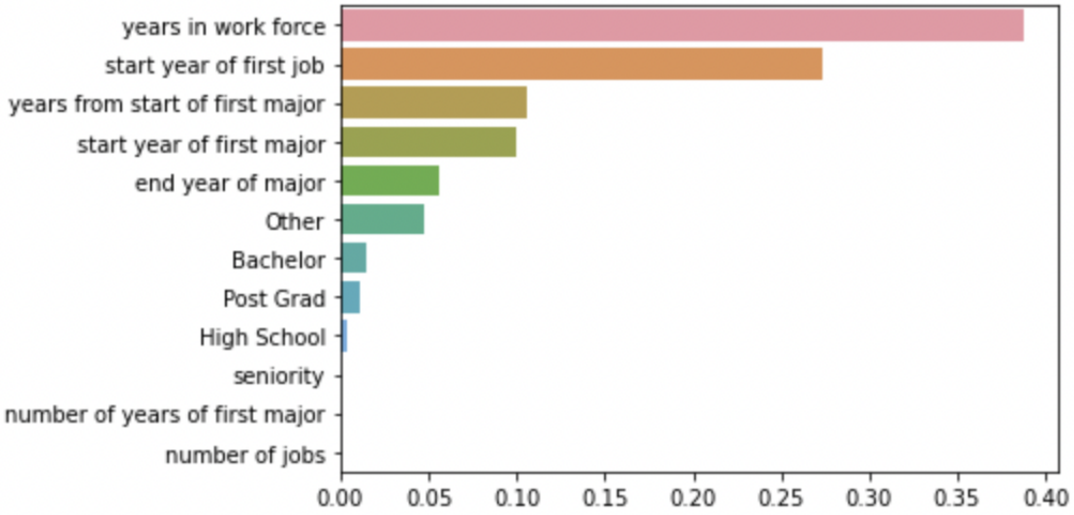
- A similar relationship is observed between the target variable ‘age’ and the other features through the scatter plots. We cannot comment on the relationship between **age** and the **binary variables Bachelor, High School, Post Grad and Other**, which indicate what is the category of the user’s first major. This is because correlation cannot be used as a statistic for categorical variables.
- There is no significant relationship between **age** and **the number of jobs and the number of years of first major.**
- The data has a lot of outliers as can be observed from the boxplots. Hence the outliers are treated and better results are obtained.

Modeling

We use the following two models for training and testing on our training data. This is a regression problem so we use regression models. The columns containing dates and strings have been dropped before proceeding with the modeling.

Random Forest

- We remove the missing values from the data because imputing them would require taking the mean/median which would not be a good step at training.
- The metrics used for model evaluation are MSE and RMSE and MAE.
- The feature importance has been plotted as follows:



- This implies that the most important features are **years in the workforce, start year of first job, years from start of first major** and there can be a possible feature elimination of **number of years of first major, number of jobs.**

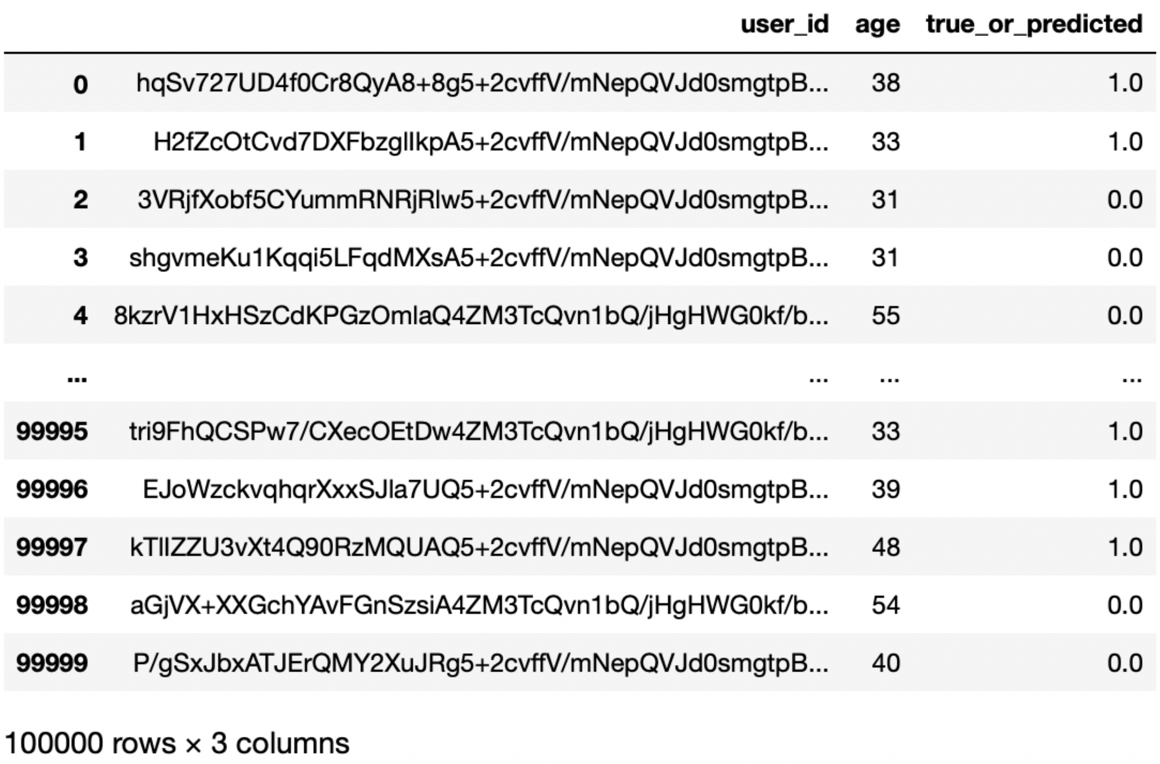
XGBoost

- The missing value data has been included in the training data for the model because XGboost is known to handle missing values in training and missing values in training data are important so that the model learns to predict well on missing values too. The original data has a lot of missing values and in cases such as when only the start year of the job or only the number of years from the first major are present the model needs to accurately predict the age.
- The metrics used for model evaluation are MSE and RMSE and MAE.
- Feature elimination has also been implemented.

Model	Training			Testing		
	MAE	MSE	RMSE	MAE	MSE	RMSE
Random Forest	0.016	0.008	0.089	0.041	0.048	0.21
XGBoost	0.10	0.026	0.16	0.15	0.089	0.29
XGBoost (Feature Elimination)	0.09	0.02	0.14	0.133	0.073	0.27

Conclusion:

We observe the metrics are similar for all the models but we predict the final values with XGBoost with feature elimination because it trains on missing values and predicts missing values that have been included in the training datasets. The XGBoost model with feature elimination has a better Mean MAE as can be seen from the code. Hyper parameter tuning could be a potential next step for better results. The data for prediction for all the users is prepared in the similar way as above and the age has been predicted for all the users.
The output data frame is as follows:



	user_id	age	true_or_predicted
0	hqSv727UD4f0Cr8QyA8+8g5+2cvffV/mNepQVJd0smgtpB...	38	1.0
1	H2fZcOtCvd7DXFbzgIIkpA5+2cvffV/mNepQVJd0smgtpB...	33	1.0
2	3VRjfXobf5CYummRNRjRlw5+2cvffV/mNepQVJd0smgtpB...	31	0.0
3	shgvmeKu1Kqqi5LFqdMXsA5+2cvffV/mNepQVJd0smgtpB...	31	0.0
4	8kzrV1HxHSzCdKPGzOmlaQ4ZM3TcQvn1bQ/jHgHWG0kf/b...	55	0.0
...	...	...	...
99995	tri9FhQCSPw7/CXecOEtDw4ZM3TcQvn1bQ/jHgHWG0kf/b...	33	1.0
99996	EJoWzckvqhqrXxxSJla7UQ5+2cvffV/mNepQVJd0smgtpB...	39	1.0
99997	kTIlZZU3vXt4Q90RzMQUAQ5+2cvffV/mNepQVJd0smgtpB...	48	1.0
99998	aGjVX+XXGchYAvFGnSzsiA4ZM3TcQvn1bQ/jHgHWG0kf/b...	54	0.0
99999	P/gSxJbxATJErQMY2XuJRg5+2cvffV/mNepQVJd0smgtpB...	40	0.0

100000 rows × 3 columns

Shambhavi Sachin Rege, NYU