GEORGETOWN UNIVERSITY
School of Continuing Studies
Center for Continuing and Professional Education

Stephanie R. Miller | Makafui Kalefe | Molly Morrison | Lourdes Peña | Maria Hernandez

Marriageability Team

Final Project Report

Data Science Certificate | Cohort 15

Abstract

Understanding the determinants of marriage is as important as the landscape of relationships in America has shifted dramatically in recent decades. This project used supervised machine learning algorithms to predict the likelihood that a person will get married while considering factors including education, gender, age, income, and location. The TEAM encountered limitations pertaining to data and technical issues. After modeling, cross validation and grid search, the best performing model was Logistic Regression. Among other things, future work could include the creation of an app to predict marriageability.

Introduction

The landscape of relationships in America has shifted dramatically in recent decades. According to a 2010 Pew Research Center nationwide survey, trends of the past 50 years have led to a sharp decline in marriage and a proliferation of novel family structures. In 1960, two-thirds (68%) of all twenty-somethings were married. In 2008, just 26% were. As overall rates of marriage decrease, the number of adults cohabitating with a partner is on the rise (Raley, Sweeny and Wondra, 2015). Shifts in marriage patterns vary greatly based on class, age, and race (Taylor, 2010). Although social scientists debate whether today's young people will eventually marry in the same numbers as earlier generations, marriage remains commonplace (Raley, Sweeney and Wondra, 2015).

<u>Factors that Influence Marriage</u>

There is extensive empirical literature about the determinants of marriage, including the importance of economic factors, such as occupation and income, as well as demographic, socio-cultural, and psychological factors (Alm and Whittington, 1999; Altonji and Vidangos, 2014).

Age. Today, young adults in the United States are waiting longer to marry than at any other time in the past century (U.S. Census Bureau). The median age of marriage has risen to 30 for men and 28 for women in 2019, up from 23 for men and 20.8 for women in 1970 (U.S. Census Bureau). Some researchers argue that putting off marriage is due to a shift in values, personal goals, and roles that differ from previous generations. For example, despite being known as the generation that gave rise to the "hookup culture" through apps like Tinder, research finds that millennials are more likely to prioritize career goals, financial stability, and finding someone who shares their values before settling down (Rabin, 2018).

Race. There are also important racial and ethnic differences in the changing marriage patterns. For example, compared to both white and Hispanic women, black women marry later in life and are less likely to marry at all (U.S. Census Bureau). Data from U.S. Census Bureau's American Community Survey for 2008–12 indicated that nearly nine out of 10 white and Asian/Pacific Islander women had ever been married by their early 40s, as had more than eight in 10 Hispanic women and more than three-quarters of American Indian/Native Alaskan women; yet fewer than two-thirds of black women reported having married at least once by the same age.

While social scientists can't fully account for the racial and ethnic differences in marriage, research best supports explanations that involve labor market disparities and other structural disadvantages that disproportionately affect marginalized identities, such as rising incarceration rates and inequities in the education system (Raley, Sweeney and Wonda, 2015).

Education. Individuals with higher levels of education are more likely to get married, whereas individuals with a high school diploma or less are less likely to get married (Raley, Sweeney and Wonda, 2015). Compared to their more highly educated counterparts, people without a college degree are less likely to achieve the economic security thought to be necessary for marriage, and those who do marry are more likely to divorce (Raley, Sweeney and Wonda, 2015). Research also finds that women tend to marry partners who have accumulated at least as much schooling as they have (Raley, Sweeney and Wonda, 2015). Interestingly, partners who meet through family have lower levels of education than partners who meet through other intermediaries (Falcon, 2015).

Income. Literature on the relationship between marriage and income is somewhat contradictory (Alm and Whittington, 1999): a higher earning capacity makes one a more attractive spouse, but it makes one more independent and less likely to have a financial need for marriage. For example, T. P. Schultz (1994) found that for white women, better wage opportunities led to a lower probability of marriage, whereas Keeley (1979) found that higher income leads to earlier marriage and an overall higher probability of ever marrying. One study found that for same-sex couples, equal earnings reduces likelihood of breakup, while equal earnings increases the likelihood of breakup for heterosexual couples (Weisshaar, 2014).

Employment. Individuals in certain professions and industries are more likely to be drawn to each other, according to a Bloomberg analysis of 2014 American Community Survey. For example, male firefighters most often marry female nurses, while female nurses most often marry managers. The most common marriage is between grade-school teachers. The analysis also found high-earning women (e.g. doctors, lawyers) tend to pair up with their economic equals, while high-earning men pair up with individuals across the earnings spectrum.

Location. Rural residents may face different marriage and labor markets from urban residents (Alm and Whittington, 1999). Prior research suggests that individuals in southern states and rural areas are more likely to get married at a younger age compared to individuals living in urban areas (Bramlett & Mosher, 2002; Goldscheider & Waite; McLaughlin, Lichter, & Johnson, 1993).

Project Overview

The purpose of this project is to predict marriageability, or the chance a person will be married based on selected characteristics. Specifically, this project aims to address two research questions: 1) Can we predict marriageability based on key demographic characteristics, economic characteristics, and geographic location? and 2) What factors are related to marriageability? We hypothesize that marriageability (or the likelihood that an individual will be married) can be predicted when factors such as race, sex, education, income, occupation sector, and location are considered. Using the data science pipeline, this project explores data related to marriage in the United States and uses supervised machine learning algorithms to make predictions (Figure 1).

Project Pipeline

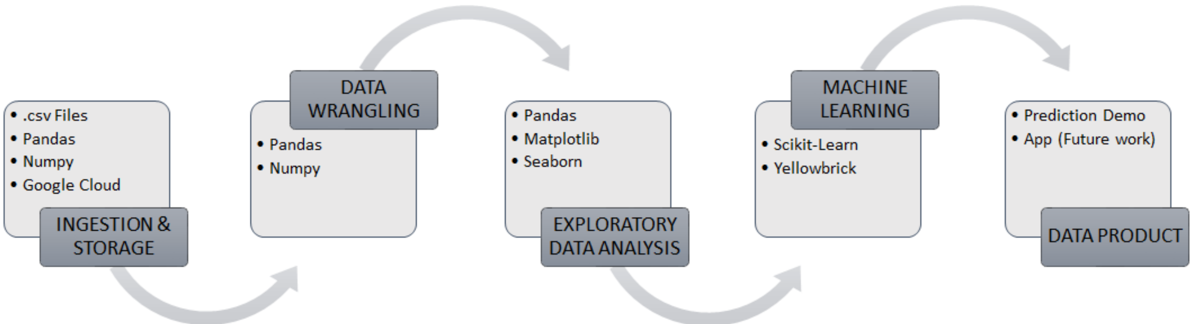


Figure 1. Marriageability Data Science Pipeline.

Ingestion: Data & Storage

Data for this project are from the American Community Survey (ACS), a yearly survey administered by the U.S. Census Bureau. The ACS is part of the decennial census, serving as an intermediary for the long form that is sent to U.S. households every 10 years. Since 2005, data for the ACS has been collected using four sequential methods: paper questionnaires sent through the mail, phone interviews, personal, one-on-one visits by Census interviewers, and internet.[1] The Census Bureau releases tabulated and untabulated ACS data.

ACS data is available since 1996 in 1-year, 3-year or 5-years Public Use Microdata Sample (PUMS) data files.[2] For each survey year, the PUMS includes separate population and housing unit record data files that includes information on marital status, age, sex, race, employment status, occupation, income, access to internet, living quarters, mortgage costs, etc. A full list of data collected via the ACS can be found here.

Each record in the person-level file represents a single person, or, in the household-level dataset, a single housing unit. In the person-level file, individuals are organized into households, making possible the study of people within the context of their families and other household members.

For this project, we utilized the 2017 ACS 1-year person- and household-level datafiles, downloaded from the online PUMS site.[3] The files were stored using a shared Google Cloud site. The 2017 ACS 1-year data were collected between January 1, 2017 and December 31, 2017 and includes data for areas with populations of 65,000 or more. The data includes information on approximately 1% of the U.S. population. The original data file includes 3,190,040 person-level instances and 1,392,399 unique household-level instances.

[1] The internet option was added in 2013 to simplify the collection and reduce costs. Beginning in 2017, the U.S. Census Bureau discontinued phone interviews as a method to follow-up with non-respondents.

[2] The U.S. Census Bureau discontinued the ACS 3-year estimates. The last set of 3-year datafiles were produced for 2011-2013 survey years.

[3] Four .csv files were downloaded, two per file type. The person datafiles included: psam_pusa.csv and psam_pusb.csv. The household datafiles included: psam_husa.csv and psam_husb.csv.

Wrangling

The U.S. Census Bureau provides users with a wealth of documentation, including a comprehensive data dictionary. The data for this project required minimal wrangling. However, the final analytic dataset was reduced to include only those persons who were at least 18 years of age.[4] The final data set includes 2,530,726 instances and 93 features (Appendix A).

Missing Data

Missing data was not a concern for this project. There were instances of blank cells in the household datafiles. However, blank cells in Census datafiles denote "not applicable" rather than missing. Therefore, in the final dataset for this project, blanks cells were set to zero (0).

Target

The target for the project is a binary classifier, where 0 = *Not Married* and 1 = *Married*. *Not Married* includes instances where an individual indicated that they were single, divorced, or widowed. *Married* includes instances where an individual indicated that they were married or separated. Nearly 45% of the instances were classified as *Not Married*, and 55% were classified as *Married*, which means the target is relatively well balanced (Table 1).

Table 1. Distribution of Target Classifier for Marriageability Project

Target Classifier	**n**	**%**
Married (1)	1,404,644	55.5
Not Married (0)	1,126,082	44.5
TOTAL	2,530,726	100

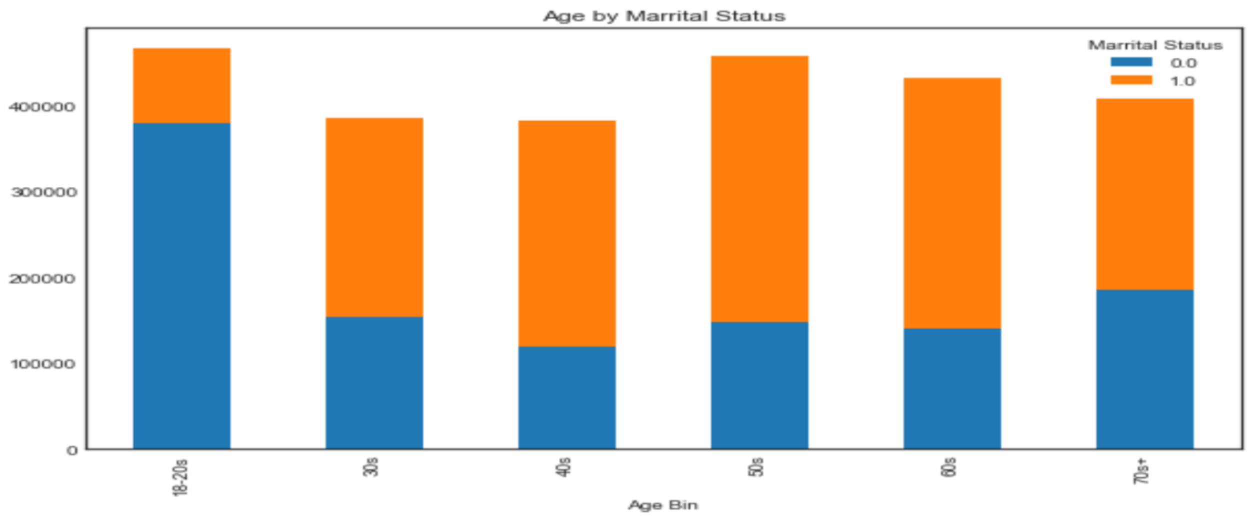
Features

Based on research and expert knowledge, the final features selected for modeling include information on location, employment, socialization, educational attainment level, income, age, sex, citizenship status, race/ethnicity, disability status, language, and the presences children. All binary features were one-hot encoded to 1/0.

The age and income features were transformed into bins. The average age of persons in the dataset is 49 years (min = 18 years, max = 96 years). Age was transformed into 6 bins: [18 - 29 years], [30 - 39 years], [40 - 49 years], [50 - 59 years], [60 - 69 years], and [70 years or older] (Figure 2).

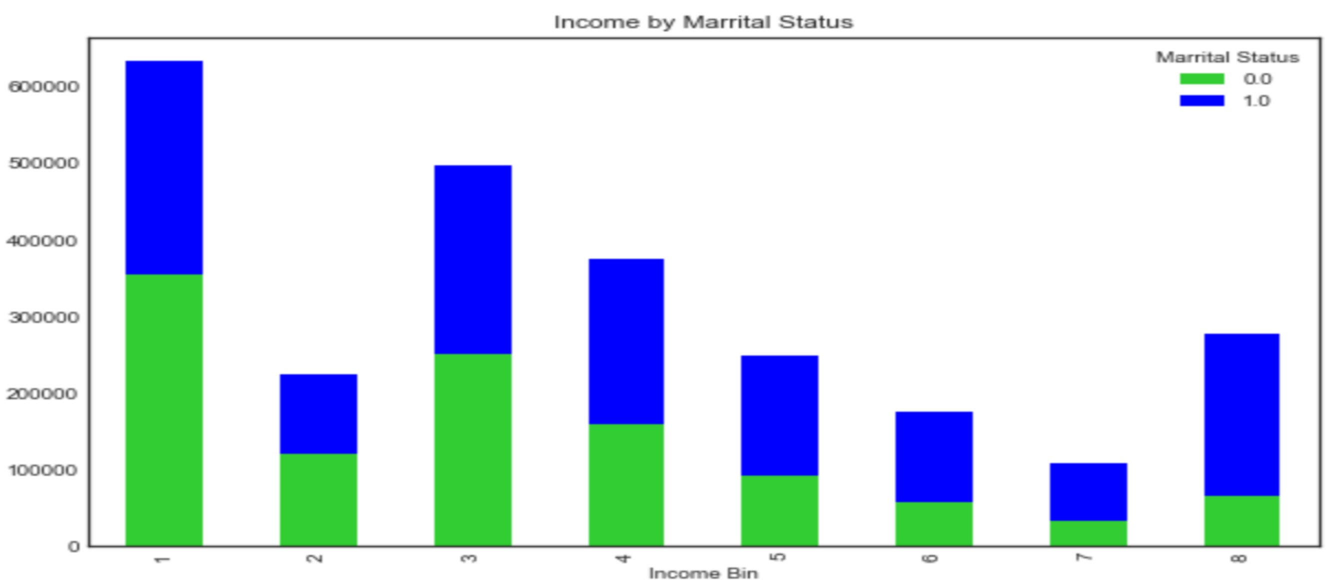
[4] The minimum statutory marrying age varies by state in the United States. Alaska and North Carolina have the lowest legal marrying age, 14 years. Some states with legal marrying ages that are less than 18 years require parental and/or judicial consent.

Figure 2. Age by Marital Status.



The average income of persons in the dataset is $43,2015 (min = 0, max = $1,580,488).[5] Income was transformed into 8 bins: [0 - 9,999], [10,000 - 14,999], [15,000 - 29,999], [30,000 - 44,999], [45,000 - 59,999], [60,000 - 74,999], [75,000 - 89,999], [90,000 or more] (Figure 3).

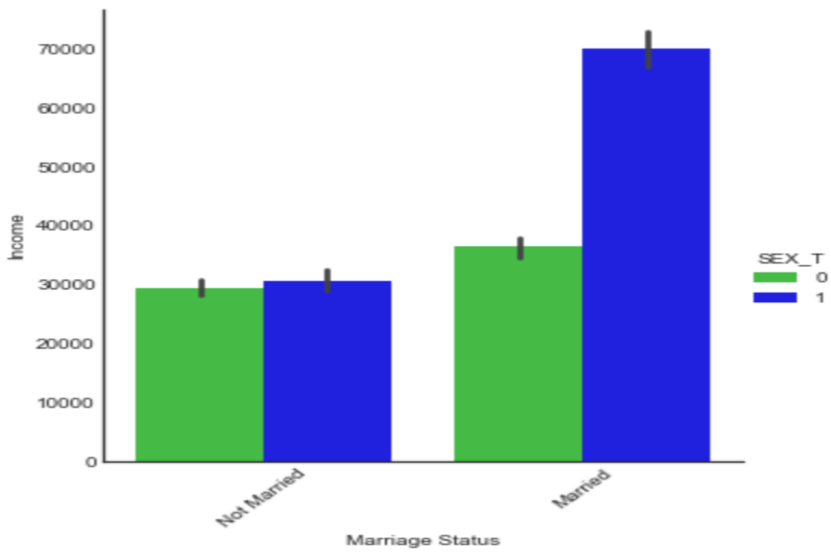
Figure 3. Income by Marital Status.



[5] Income was adjusted for inflation.

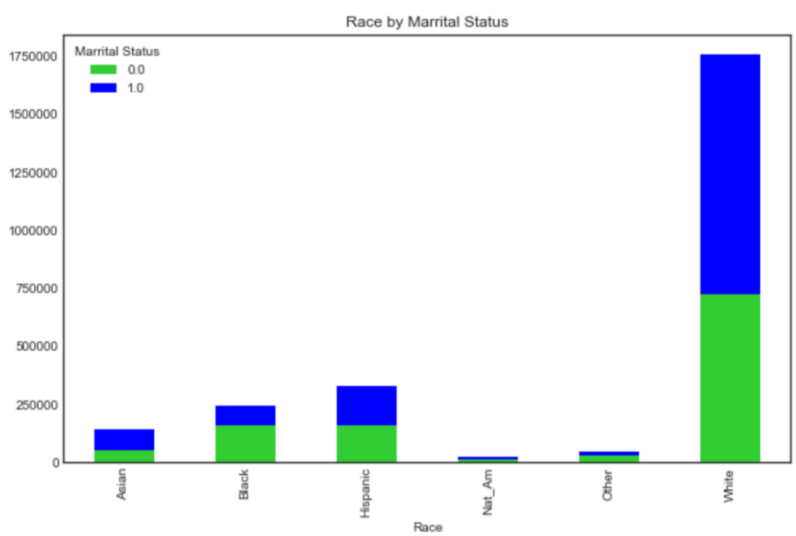
Data also shows that older individuals are more likely to be married, while younger individuals between 18 to 29 years of age are less likely to be married (Figure 2). In terms of income, married men earn on average more than double the income of married women ($70,751 compared to $35,415; Figure 4).

Figure 4. Marriage by Income and Sex



An analysis of race and marital status in the data showed that there are more married White people in the data than any other race/ethnicity (Figure 5). It is important to note that race classes are significantly unbalanced, potentially introducing bias in our model. In this dataset, nearly 70% of people are White. Latinos account for 13% of the data, Blacks account for nearly 10% of the data. Collectively, Asians, Native Americans and Other races account for less than 9% of the data.

Figure 5. Marriage by Race



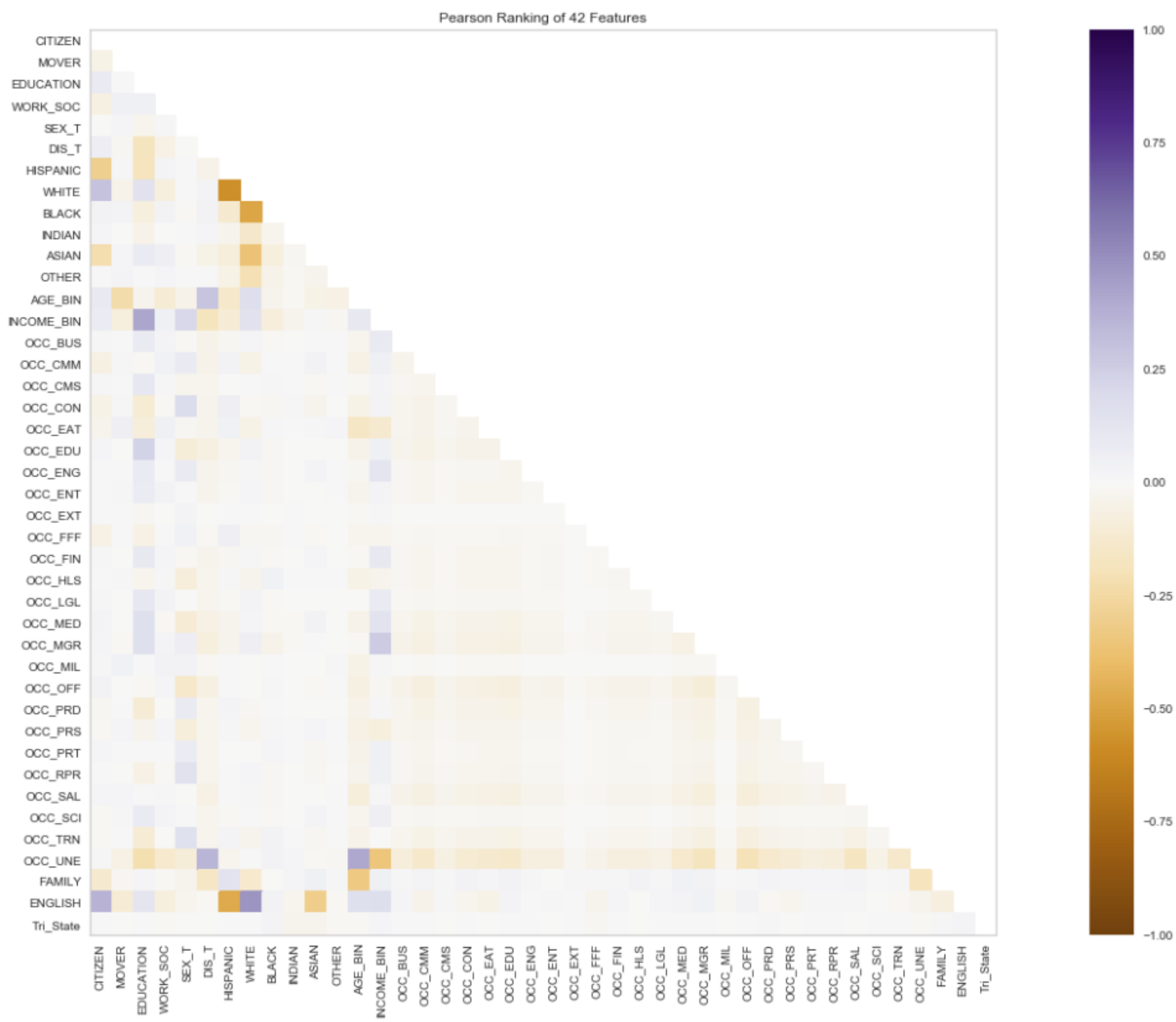
Location Feature

For this project, location is an important aspect of the model. The dataset includes 2010 Public Use Microdata Areas (PUMA), which are geographic units used by the U.S. Census Bureau to provide statistical and demographic information for sub-state areas. PUMAs typically do not overlap and are contained within a single state. The 2010 PUMAs include 2,378 statistical geographic areas covering the United States.

Unfortunately, given computing power and model runtime constraints, the project TEAM could not proceed with the PUMAs feature. As an alternative to PUMAs, the team proceeded with using state (ST) to denote location. The TEAM also created a feature, *Tri_State*, to denote those persons that lived in states connected by economy and geography. Those states designated as *tri--state* include Connecticut, Delaware, District of Columbia, Illinois, Indiana, Kentucky, Maryland, New Jersey, New York, Ohio, Pennsylvania, Virginia, and West Virginia. Nearly 22% of the instances for the project dataset live in a designated tri-state state.

To proceed with modeling, the TEAM analyzed feature correlation. The Pearson Ranking correlation matrix shows that there are no concerns for potential covariance (Figure 6).

Figure 6. Correlation Matrix of Features for Marriageability Project



Modeling

The objective of this project is to predict marriageability for a given person using supervised machine learning algorithms. Specifically, this is a classification problem, where the goal is to predict a class label (either married or not married). Five different classification algorithms were chosen for initial modeling, including Logistic Regression, K-Nearest Neighbor, Gaussian Naive Bayes, Gradient Boost, and Random Forest. The data was split into separate training (n = 2,024,580; 80%) and testing (n = 506,146; 20%) datasets.

For the analysis, the classifier ("MARRIED") was fit using the training dataset to evaluate and make predictions on the testing dataset. For each classification algorithm, an initial model was run, then cross validation was performed to evaluate performance using several splits of the data. Table 2 shows the performance of each initial classification algorithm. The results of the analysis that the Logistic Regression and the Gradient Boost algorithms yield the best combination of f1, precision scores. Note that the TEAM was unable to run KNN (or two additional models) given lack of computing power.

Cross-Validation

After cross-validation, the results showed that the Logistic Regression algorithm is 68% accurate on average. The Gradient Boost algorithm is 71% accurate on average. Both algorithms were produced by the six-fold cross-validation and show relatively low variance in the accuracy between folds (see Appendix 1).

Table 2. The performance of classification algorithms for Marriageability Project

Classification Algorithm	**f1**	**Precision***	**Recall***	**Average CV Score**
Logistic Regression	0.728	0.682	0.781	0.677
K-Nearest Neighbor	—	—	—	—
Gaussian Naïve Bayes	0.650	0.660	0.640	0.618
Gradient Boost	0.766	0.696	0.852	0.713
Random Forest	0.725	0.624	0.956	0.656**

*Precision and Recall are reported for "Married = 1"

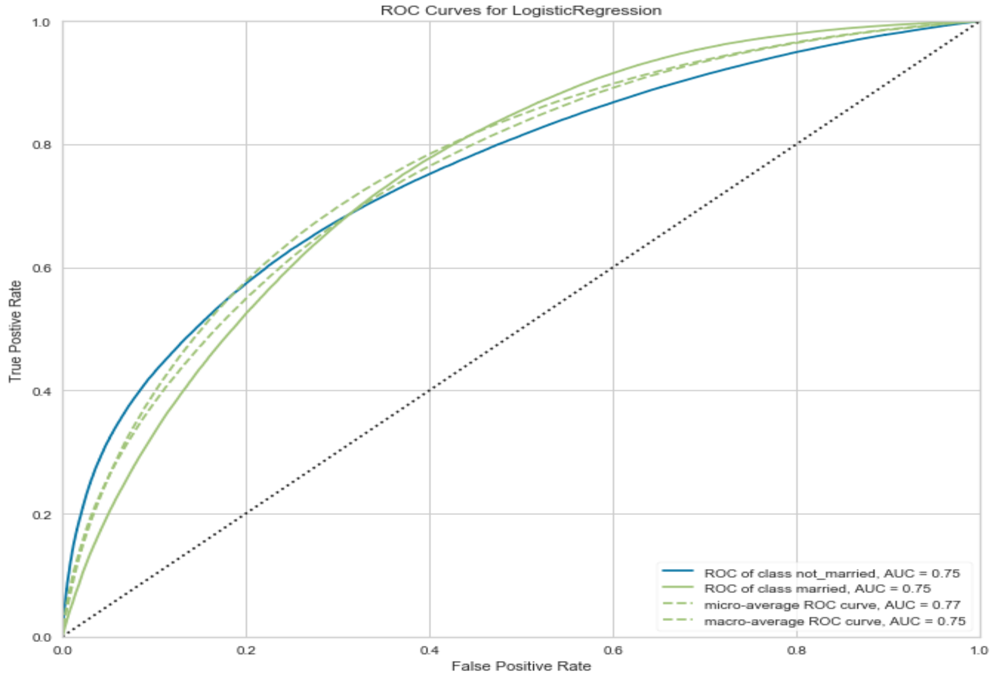
**The observed scores for Random Forest showed a relatively high variance in the accuracy between folds, ranging from 0.628 to 0.678

Hyperparameter tuning

With technical difficulties to run most models, the TEAM decided to focus on Logistic Regression only to have a minimum viable product. The next step in the process was to perform a grid search to ascertain the best hyperparameters: {'C': 0.001, 'penalty': 'l2'}, which implies a stronger l2 regularization for the model. The result was a slightly higher f1 score (0.730).

To evaluate the Logistic Regression algorithm, the TEAM analyzed the ROCAUC Curve (Figure 7). Ideally, the curve would be closer to the top left, but both ROC and AUC measures are relatively high.

Figure 7. ROC-AUC Curve for Logistic Regression



Predictions

After tuning the model with the best parameters, the TEAM tested the Logistic Regression and Gradient Boost models under different scenarios. Table 3 compares predictions for each project TEAM member members.

Table 3. Predictions - Model output for Linear Regression and Gradient Boost Classifier

	LR		GBC	
	Not Married	Married	Not Married	Married
Lulu	[0.38254617,	0.61745383]	[0.35778459,	0.64221541]
Stephanie	[0.53220923,	0.46779077]	[0.58407321,	0.41592679]
Maria	[0.27268135,	0.72731865]	[0.34414361,	0.65585639]
Molly	[0.50298999,	0.49701001]	[0.78000214,	0.21999786]
Makafui	[0.56450502,	0.43549498]	[0.47930086,	0.52069914]

Insights and Other Considerations

Data Limitations

The data source used for this project was limited in that it did not include import features such as sexual orientation and adults cohabitating with a partner. According to the Pew Research Center, in 2019 the number of Americans living with an unmarried partner increased 29% since 2007 and is rising the quickest among Americans ages 50 and older (Geiger and Livingston, 2019).

Data Bias

As discussed in the previous section, 70% of people in the data for this project were White, and aside from Latinos who accounted for 13%, other race classes represented less than 10% of the data for this project. The underrepresentation of non-White classes in this project is a concern, and we acknowledge that the models used to predict marriageability may be biased towards the majority class. Future work for this project should consider equality and fairness to ensure that the same *opportunity* is afforded to every instance in the model. Some of the strategies to mitigate bias in the project are sampling-out a proportion of the dominant classes to create balance in the data or using algorithms that learn from bias using dimensional reduction. Another strategy under consideration is to incorporate older years of ACS data, specifically pulling data for the underrepresented race classes.

Project Challenges

The biggest challenge in completing this project was run-time. The data size interfered with the performance and efficiently in the use of Jupyter notebooks. The TEAM experienced this issue when performing complex data transformations and when running the machine learning algorithms. For future projects, the TEAM will consider parallel processing techniques and the use of cloud servers to improve processing performance.

Future Work

Given the project’s limitations, the next steps involve using a more inclusive dataset and expanding the target to include a broader definition of relationships, such as common law marriage and domestic partnership. Future work also involves exploring additional features that may be impact likelihood of marriage such as the use of dating apps, urban vs. rural locations, and the socialness of a city (i.e. number of bars, public parks, etc.). The goal of this project is to create a user-friendly, web-based app that allows users to input both their personal characteristics as well as selecting what features are most important in their partner. The app would then return the individual’s marriageability score and recommend three alternative locations where their score registers higher based on inputs and preferences.

References

Alm, J. and Whittington, L.A. 1999. "For love or money? The impact of income taxes on marriage. Economica (1999) 66, 297-316.

Altonji, J. and Vidangos, I. 2014. "Marriage dynamics, earnings dynamics, and lifetime family income." Working Paper. Yale University.

Becker, G. S. 1973. A theory of marriage: Part I. Journal of Political Economy, 81, 813-46.

Becker, G.S. 1974. A theory of marriage: Part II. Journal of Political Economy, 82 (2, Part II), SI 1-S26.

Bramlett MD, Mosher WD. Vital and Health Statistics. 22. Vol. 23. National Center for Health Statistics; Hyattsville, MD: 2002. Cohabitation, marriage, divorce, and remarriage in the United States.

Bruze, G., Svarer, M., and Weiss, Y. "The dynamics of marriage and divorce." *Journal of Labor Economics*, Vol. 33, No. 1 (January 2015), pp. 123-170.

Falcon, M. 2015. "Family influences on mate selection: Outcomes for homogamy and same-sex coupling." Stanford University. Working Paper.

Geiger, A.W. and Livingston, G. "8 facts about love and marriage in America." *Fact Tank: News in the Numbers*, Pew Research Center, 13 February 2019. https://www.pewresearch.org/fact-tank/2019/02/13/8-facts-about-love-and-marriage/

Glick, P.C. and Norton, A.J. 1971. "Frequency, duration and probability of marriage and divorce." *Journal of Marriage and Family,* Vol. 33, No. 2, Decade Review. Part 3 (May 1971), pp. 307-317.

Goldscheider FK, Goldscheider C. Leaving home before marriage: Ethnicity, familism, and generational relationships. University of Wisconsin Press; Madison, WI: 1993.

Gould, E. 2008. "Marriage and career: the dynamic decisions of young men." *Journal of Human Capital*, 2008, vol. 2, no.4.

HHS. 2003. "The determinants of marriage and cohabitation among disadvantaged Americans: research findings and needs." Marriage and Family Formation Data Analysis Project Final Report. Prepared by Abt Associates Inc.

Lefgren, L. and Mcintyre, F. 2006. "The relationship between women's education and marriage outcomes." *Journal of Labor Economics,* Vol. 24, No. 4 (October 2006), pp. 787-830.

Nugent, C.N. and Daugherty, J. 2018. "A Demographic, Attitudinal, and Behavioral Profile of Cohabiting Adults in the United States, 2011–2015." National Health Statistics Reports. Number 111 May 31, 2018.

Pew Research Center. 2010. "Marrying Out. One-in-seven new U.S. marriages is interracial or interethnic."

Pew Research Center. 2010. "The Decline of Marriage and Rise of New Families."

Rabin, R. 2018. Put a Ring on It? Millennial Couples Are in No Hurry. *New York Times.*

Raley, R.K., Sweeney, M. and Wondra, D. 2015. “The growing racial and ethnic divide in U.S. marriage patterns.” Author manuscript published in final edited form as: *Future Child.* 2015; 25(2): 89-109.

Taylor, P., 2010. The Decline of Marriage and Rise of New Families. Pew research center.

Weisshaar, K. 2014. “Earnings equality and relationship stability for same-sex and heterosexual couples.” Stanford University.

Appendix A. Features Included in the Machine Learning Algorithms for the Marriageability Project

ACS VARIABLE	DEFINITION	PROJECT FEATRUE	LOGIC MODEL
CIT	Citizenship status	**CITIZEN** 1 = Yes 0 = No	Demographic
ST	1=Alabama/AL 2=Alaska/AK 4=Arizona/AZ 5=Arkansas/AR 6=California/CA 8=Colorado/CO 9=Connecticut/CT 10=Delaware/DE 11=District of Columbia/DC 12=Florida/FL 13=Georgia/GA 15=Hawaii/HI 16=Idaho/ID 17=Illinois/IL 18=Indiana/IN 19=Iowa/IA 20=Kansas/KS 21=Kentucky/KY 22=Louisiana/LA 23=Maine/ME 24=Maryland/MD 25=Massachusetts/MA 26=Michigan/MI	*One-hot encoded*	Location

	27=Minnesota/MN 28=Mississippi/MS 29=Missouri/MO 30=Montana/MT 31=Nebraska/NE 32=Nevada/NV 33=New Hampshire/NH 34=New Jersey/NJ 35=New Mexico/NM 36=New York/NY 37=North Carolina/NC 38=North Dakota/ND 39=Ohio/OH 40=Oklahoma/OK 41=Oregon/OR 42=Pennsylvania/PA 44=Rhode Island/RI 45=South Carolina/SC 46=South Dakota/SD 47=Tennessee/TN 48=Texas/TX 49=Utah/UT 50=Vermont/VT 51=Virginia/VA 53=Washington/WA 54=West Virginia/WV 55=Wisconsin/WI 56=Wyoming/WY		
N/A		**Tri_State** 0 = Does not live in a tri-state state	Location

		1 = Lives in a tri-state state (Connecticut, Delaware, District of Columbia, Illinois, Indiana, Kentucky, Maryland, New Jersey, New York, Ohio, Pennsylvania, Virginia, and West Virginia)	
AGEP	Age	**AGE_BIN** 1 = 18 – 29 2 = 30 – 39 3 = 40 – 49 4 = 50 – 59 5 = 60 – 69 6 = 70+	Demographic
JWTR	Means of transportation to work	**WORK_SOC** 0 = No contact while travelling to work or does not work 1 = Contact with people while travelling to work	Social
MAR	Marital status	**MARRIED** 0 = Not married (single/never married, divorced and widowed) 1 = Married (married, separated)	Outcome
MIG	Mobility status (time lived in house)	**MOVER** 0 = Did not move or change location 1 = Moved, changed location	Social
SCHL	Educational attainment	**EDUCATION** 0 = No HS dip./GED 1 = HS diploma or GED 2 = < college degree/AA 3 = bachelor's degree 4 = graduate degree	Demographic
SEX	*Sex* 1 = Male 2 = Female	**SEX_T** 0 = Female 1 = Male	Demographic
PINCP	Total person's income	**INCOME_BIN** **Adjusted using ADJINC*	Economic

		1 = 0 - 9,999 2 = 10,000 - 14,999 3 = 15,000 - 29,999 4 = 30,000 - 44,999 5 = 45,000 - 59,999 6 = 60,000 - 74,999 7 = 75,000 - 89,999 8 = 90,000+	
RAC1P		*One-hot encoded* HISPANIC WHITE BLACK INDIAN ASIAN OTHER	Demographic
OCCP	Occupation sector	*One-hot encoded* MGR BUS FIN CMM ENG SCI CMS LGL EDU ENT CMM MED HLS PRT EAT CMM PRS SAL OFF	Economic

		FFF CON EXT RPR PRD TRN MIL *UNE - Unemployed	
DIS	*With/without disability* 1 = With a disability 2 = Without a disability	**DIS_T** 0 = No disability 1 = Disability	Demographic
HUPARC	*Presence of related children* 1 = Presence of related children under 6 years only 2 = Presence of related children 6 to 17 years only 3 = Presence of related children under 6 years and 6 to 17 years 4= No related children present	**FAMILY** 0 = No related children present 1 = Related children present	Demographic
HHL	*Household language* 1 = English only 2 = Spanish 3 = Other Indo-European languages 4 = Asian and Pacific Island languages 5 = Other language	**ENGLISH** 0 = Non-English speakers 1 = English speakers	Demographic