Team 15

Problem Statement:

In the project, we attempt to answer the question: how are tech-industry jobs affected post-recession era? Even though we observe that many places have had more job openings as the economy recovers, it is possible that different states recover at different rates, and the job markets may also depend on local industry, education, and other factors. To answer the question, we evaluate how different potential features impact the total number of job openings per month on a state level.

Data Description & Feature Selection:

We focus exclusively on technology industry. To select all data entries from technology industry, we select all companies with the "Software & IT Services" in "trc_industry_group" column from *companies* dataset. Then we choose all job openings data from *jobs* dataset with the companies we select. We measure the number of job openings by aggregating the data based on "created_date" and "last_checked_date". The time between job created date and last checked date is considered to be active job opening. Number of job openings for every US state per month is the data entry for each row in the constructed data set. We assume that job openings are affected by factors such as education level, population income level, local housing price, and financial market performances, we select the following features:

- **Level of education** is divided into four categories: before obtaining high school degree, high school degree, bachelor's degree, and graduate-level degree. There are four categories in the fields of educational study: computer science (including mathematics and statistics), business, other engineering majors, and non-technical majors. All the numbers are captured in the state level.
- **Population Income** is extracted from *econ_state* dataset, using gdp per_capita for each state as measurement for the income level of the population of the state.
- **Local Housing Price** is indicated by the aggregated "Zillow Home Value Index" for each state from *real_estate* dataset.
- **Financial Market Performance** is measured by *NASDAQ 100 Technology Sector Index* (*NDXT*) scraped from Yahoo.com on a monthly rate.

Non-Technical Executive Summary:

1. The number of technical industry job openings are strongly related to the GDP and people's education levels.
2. Job opening counts in the technical industry over time are also positively correlated with the Nasdaq 100 technology index.
3. Higher housing index value indicates that more people are moving to the area, which also suggests a stronger economy.

4. Our analyses have found that the number of job openings in a state is affected strongly by the overall economy like the stock price and the local economy, which is measured by the housing index. Furthermore, there are more job openings in the state where the more population percentage acquiring a high school diploma and even a bachelor's degree.
5. The prediction results show that there is still an increasing trend for job openings in California in 2016 and 2017.

Technical Executive Summary:

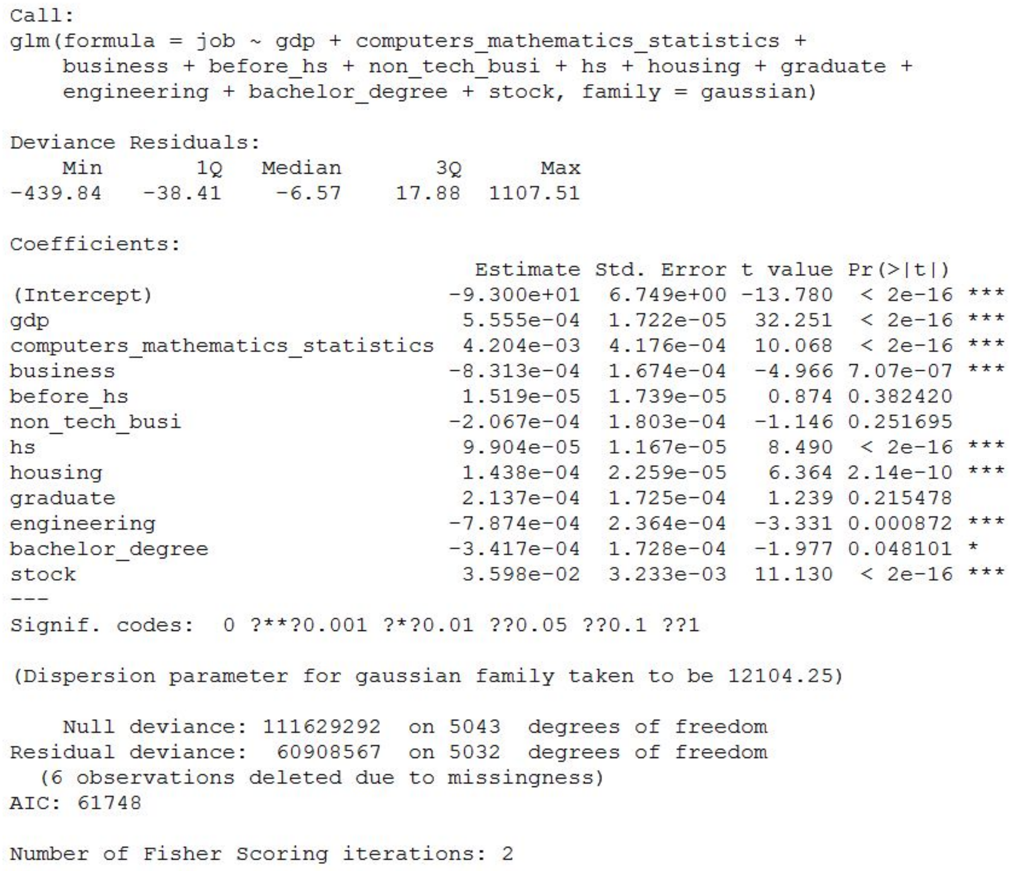
Feature Selection Model:

We train a univariate feature selection model based on *Chi-Squared* test and find the following if we reduce the number of features by half, we get five features all related to education: before-high school, high-school, bachelor, graduate and non-science degree. If we were to choose only one feature, it is high-school degree that matters the most in the number of job openings in technology industry.

For the case where no fixed number of features are given, we apply a *l1-penalized* support vector machine classification feature selection model, which selects all related features. These include all features we select, meaning in our general linear model, we will first feed all features to see their correlations.

General Linear Model:

Job ~ GDP + Computer_Science_Major + Business_Major + Non_Technical_Major + Engineering_Major + Before_Highschool_Education + High_School_Degree + Bechelor's_Degree + Graduate_Degree + Housing_Index + Stock_Price

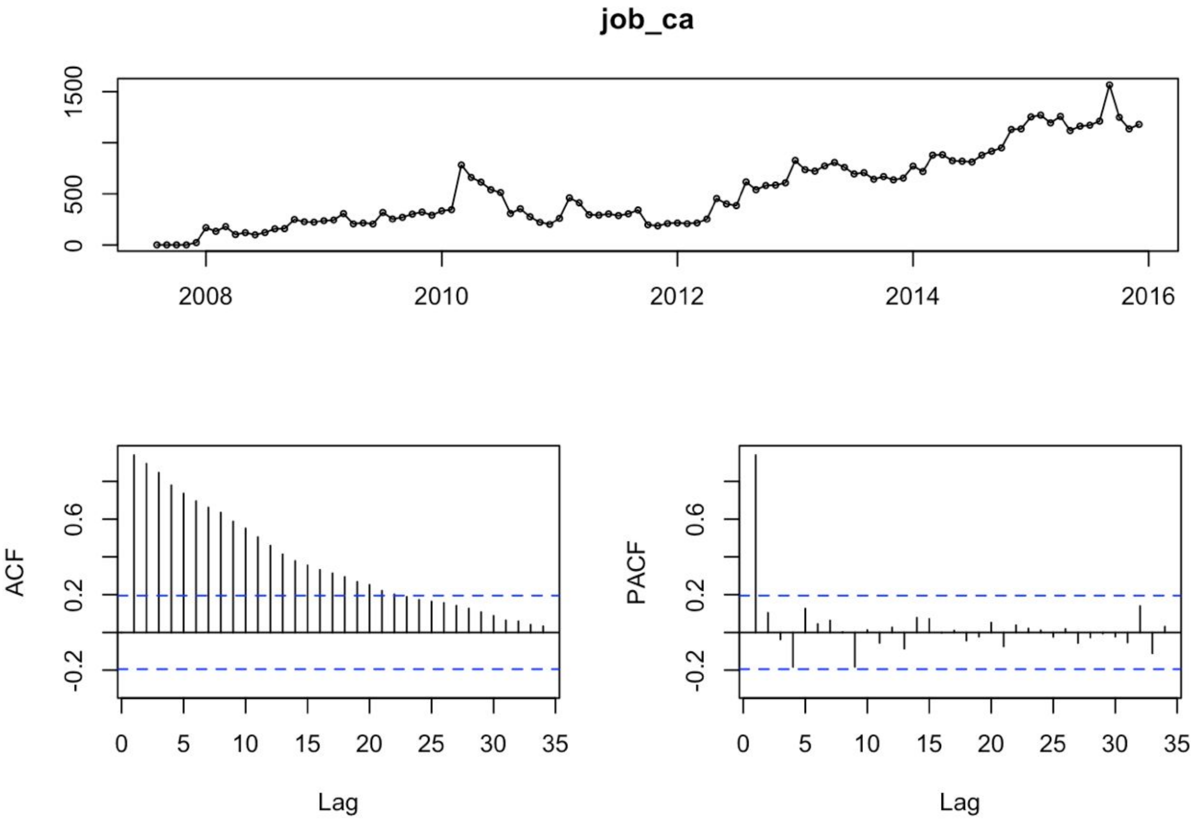


```
Call:
glm(formula = job ~ gdp + computers_mathematics_statistics +
    business + before_hs + non_tech_busi + hs + housing + graduate +
    engineering + bachelor_degree + stock, family = gaussian)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-439.84   -38.41    -6.57    17.88  1107.51

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -9.300e+01  6.749e+00 -13.780  < 2e-16 ***
gdp                               5.555e-04  1.722e-05  32.251  < 2e-16 ***
computers_mathematics_statistics  4.204e-03  4.176e-04  10.068  < 2e-16 ***
business                         -8.313e-04  1.674e-04  -4.966 7.07e-07 ***
before_hs                         1.519e-05  1.739e-05   0.874 0.382420
non_tech_busi                    -2.067e-04  1.803e-04  -1.146 0.251695
hs                                9.904e-05  1.167e-05   8.490  < 2e-16 ***
housing                           1.438e-04  2.259e-05   6.364 2.14e-10 ***
graduate                          2.137e-04  1.725e-04   1.239 0.215478
engineering                      -7.874e-04  2.364e-04  -3.331 0.000872 ***
bachelor_degree                  -3.417e-04  1.728e-04  -1.977 0.048101 *
stock                             3.598e-02  3.233e-03  11.130  < 2e-16 ***
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

(Dispersion parameter for gaussian family taken to be 12104.25)

    Null deviance: 111629292  on 5043  degrees of freedom
Residual deviance:  60908567  on 5032  degrees of freedom
  (6 observations deleted due to missingness)
AIC: 61748

Number of Fisher Scoring iterations: 2
```

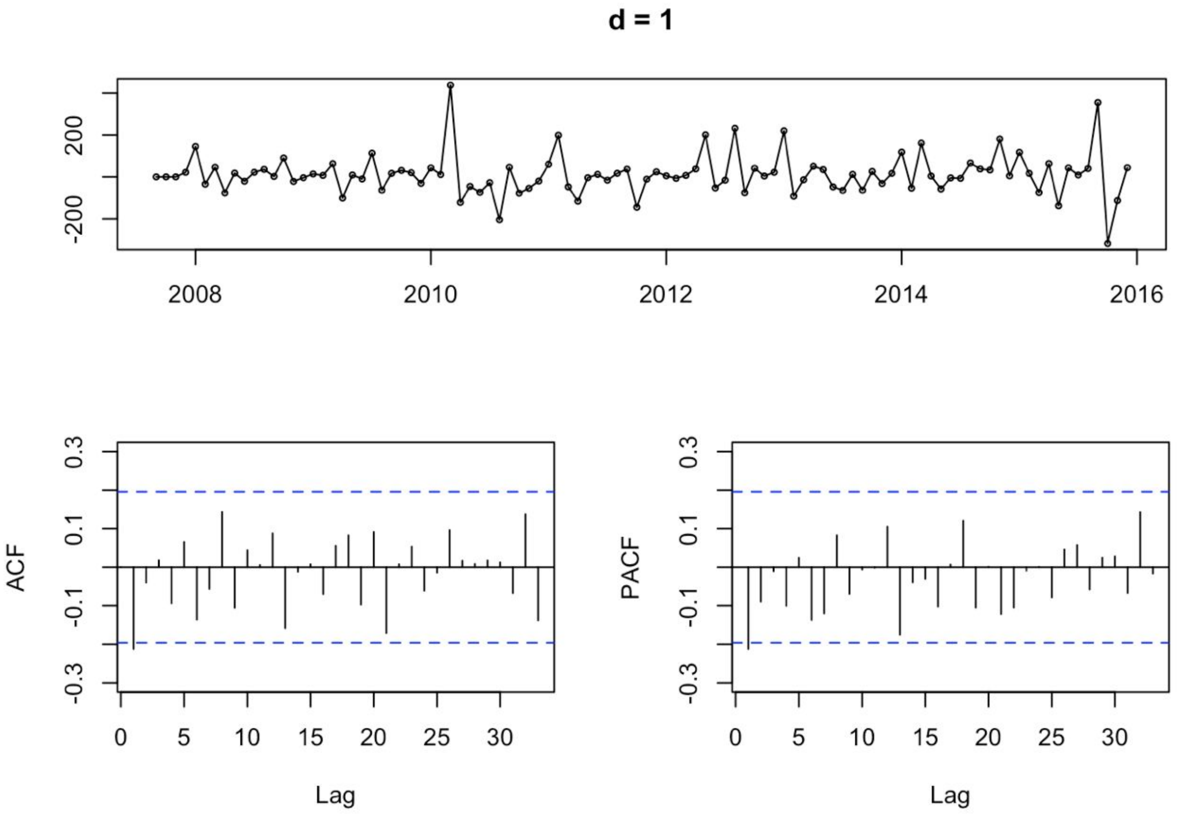
With general linear regression, the model is fitted with gaussian distributions. The results show that before graduating from high school, people with non-technical education backgrounds, and obtaining a graduate degree are less significant in predicting the job openings in technical industry. The most significant features are GDP, with computer science backgrounds, graduating from high school, housing index, and the stock market performance. This explains that the overall economy performance affects the job openings in tech industry the most. The result is also supported by the feature selection, which indicates that high school education and housing index are among the most important features. Also, since the jobs in this industry has more technical-related job openings compared to other non-tech jobs, computer, mathematics, and statistics backgrounds are more related with the number of job openings.

Time Series Model:

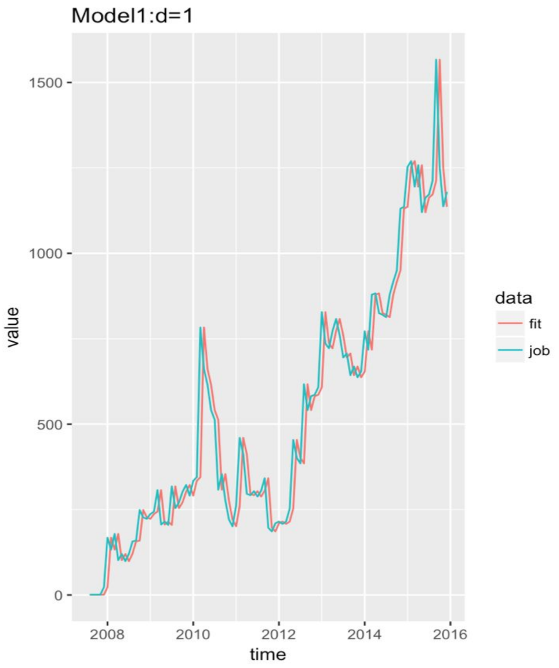
Since we aggregated the job opening data into monthly level from 2007-08 to 2015-12 within each state, we considered to fit a time series model on job opening data. Take California as an

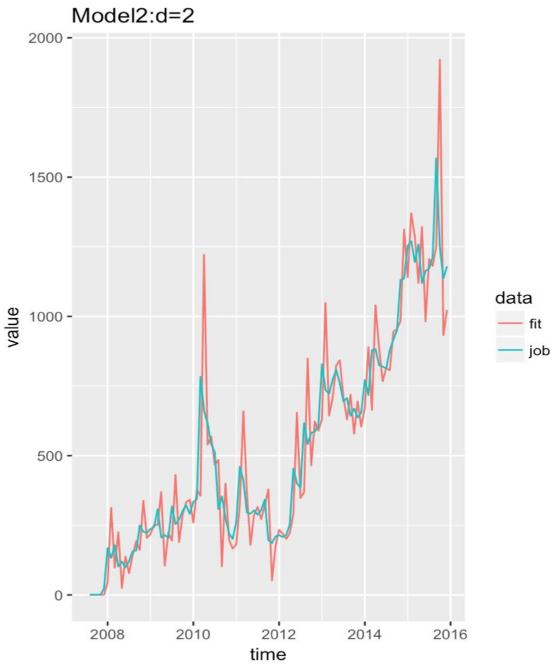


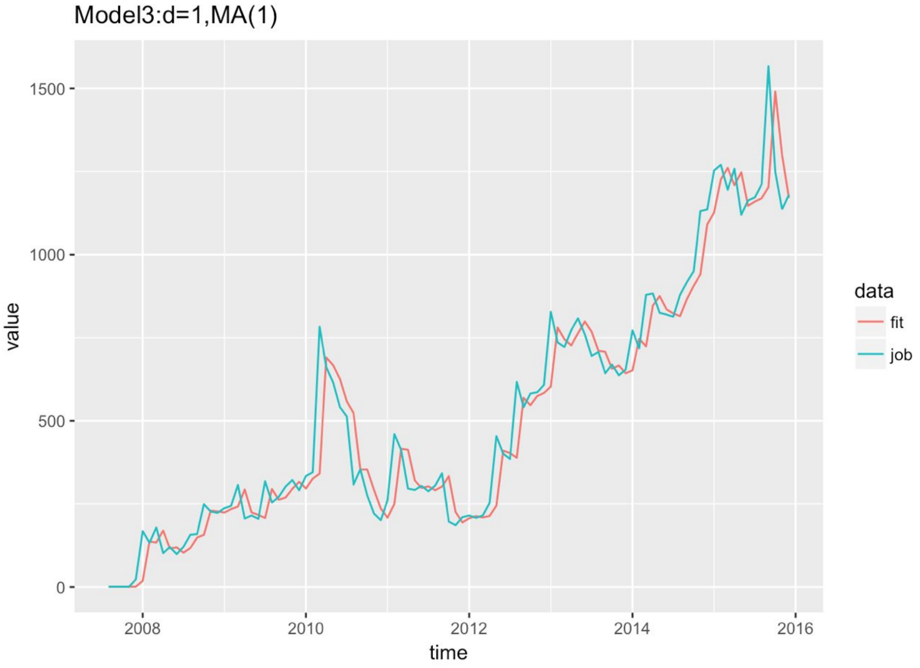
example here. First, we looked at the scatter plot, acf and pacf. The acfs are quite large in first 10 lags, then we decided to apply difference to the data, then we tried first order difference and second order difference. We found that first order difference worked better on reducing the



autocorrelation. Then we fit ARIMA(0,1,0), ARIMA(0,2,0) and ARIMA(0,1,1) models. The fitted results are displayed as belows:

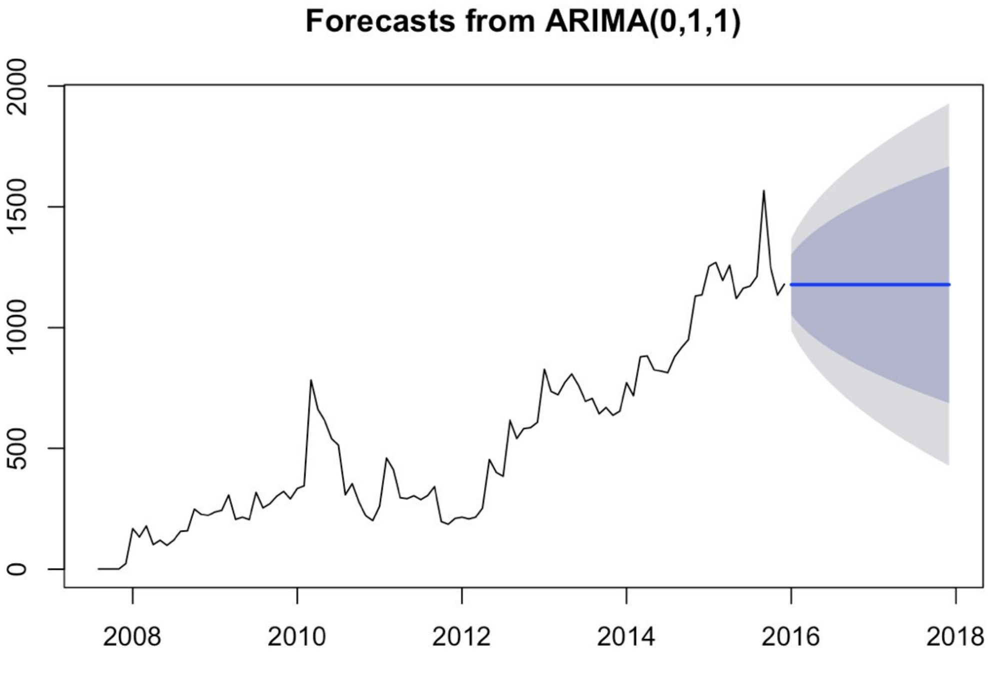






model	rmse	AICc	BIC
Arima(0,1,0)	98.69590	1205.23224241315	1207.79659627261
Arima(0,2,0)	152.54871	1280.41295570872	1282.96683844546
Arima(0,1,1)	96.62359	1203.11605989275	1208.20268892452

According to the table above, ARIMA(0,1,1) performed best as its RMSE and AIC are the smallest. We chose to predict the data with ARIMA(0,1,1) model and the visualization shows as below. The prediction results show that there is still an increasing trend for job openings in California in 2016 and 2017. The 95% confidential interval is about [750-1650].


Forecasts from ARIMA(0,1,1)