Data Analytics and Modeling: Quantitative Analysis for Economic Strategy (ECON 453)

Term Paper

Kyle Bonebrake, Minh Duong, Naman Patel

Abstract:

In this paper, we analyze survey data from over 70,000 software developers in an attempt to understand the factors influencing their salaries and attitudes towards blockchain technology. We will propose four models, with two focusing on salary determinants and two on blockchain favorability. Our study provides valuable insights into the relationship between software developer's demographic, education level, professional background, salary level and perceptions on blockchain technology. By shedding light on these associations, our findings aim to inform everyone from employees to employers.

Background:

The software engineering industry has experienced rapid growth in recent years, and as a result, there is a high demand for talented software engineers. One of the critical factors that influence the job satisfaction of a software engineer is their salary. Therefore, it is essential to understand the various factors that determine a software engineer's salary to enable companies to attract and retain top talent in the field.

Another important trend in the software engineering industry is the increasing interest in blockchain technology. With the rise of cryptocurrencies and blockchain-based applications, more and more companies are looking to hire skilled blockchain developers. However, there is still much to be learned about the attitudes of software engineers towards this emerging technology. For this reason we hope to construct a model that can predict whether or not a software developer will find blockchain technology favorable to work with.

The problem we have selected for this project is to develop accurate models that can predict a software engineer's salary and attitudes towards blockchain technology based on various factors such as location, company size, years of professional coding experience, exposure to backend technology, etc. The objective of this study is to provide insights into the factors that impact a software engineer's salary and attitudes towards blockchain technology. By developing accurate prediction models, we aim to help companies understand what factors are most important when it comes to attracting and retaining talented software engineers and identifying which software engineers may be most interested in working with blockchain technology.

Ultimately, the insights gained from this study may inform employers and help companies to develop more effective recruitment strategies to attract top talent in the industry. Moreover, this study may contribute to a better understanding of the attitudes of software engineers towards blockchain technology, a critical area of interest as blockchain-based applications continue to grow in popularity.

Data:

In this project, we choose to work with the result from StackOverflow Developer Survey 2022, which is sourced from Kaggle. This cross-sectional dataset that contains data from over 70,000 software developers responses, with over seventy attributes ranging from profile data, technology, work, community to professional information, including education level, how they learned to code, level of experience, developer roles, key territories, and demographics, to name a few. Our dataset also goes in depth about technology preferences of developers, encompassing the programs and tools they currently utilize as well as those they have expressed interest in using in the future. Additionally the data provides work information, like the size of the organization they work for, if they work remotely, and education level. The data also displayed columns for Professional development like work lifestyle, experience, years of coding both in total and professionally.

To be able to work with the dataset, we have pre-processed the data to categorize and remove outliers specific to our models. Those changes include:

- Excluding all yearly converted comp under 1,000 and over 500,000 - this was done in an effort to lower the number of outliers in our dataset.
- Categorizing companies with less than 20 employees as very small companies, and freelancers were also categorized as very small companies.
- Categorizing companies with employees between 20 and 99 as small companies.
- Categorizing companies with employees between 1000 and 4,999 as large companies
- Categorizing companies with more than 5,000 employees as very large companies.
- Standardizing individuals with "more than 50 years of experience" as 51 and standardized individuals with "less than 1 year of experience" as 0.
- Categorizing the ages from 18 – 24 years old as Young.
- Categorizing the ages from 25 – 34 years old as Normal.
- Categorizing the ages from 35 - 44 years old as Middle.
- Categorizing the ages from 45 - 54 years old as Retiring.
- Categorizing the ages from 55 - 64 years old as Senior.

Objectives/Hypothesis:

The main objective of this study is to develop models that can accurately predict the salary of a software engineer and their attitudes towards blockchain technology. Specifically, we aim to create two linear regression models that can predict a software engineer's salary based on various factors such as their country, company size, years of professional coding experience, number of technologies they have used, and whether they work remotely. We also aim to create two classification models that can predict whether a software engineer would be favorable towards blockchain technology or not. To achieve this, we will use logistic regression and Naive Bayes with various independent variables such as the number of technologies they have used, their work experiences, etc.

We hypothesize that the factors we have chosen for the linear regression models will have a significant impact on a software engineer's salary. Specifically, the hypotheses are:

- The independent variables on the salary prediction models will individually and jointly be statistically significant in determining the salary.
- The independent variables on the blockchain favor prediction model will individually and jointly be statistically significant in determining whether a person likes blockchain or not.

We will also compare the models to each other regarding their accuracy, explanatory power and predictive power to acquire more insights into factors affecting the dependent variables, and will choose the better model for our objective.

Literature Review:

Title of Research Paper: *Who Will Leave the Company?: A Large-Scale Industry Study of Developer Turnover by Mining Monthly Work Report*

This research paper examines the possibility of predicting whether a developer will leave a company within one year based on six years of monthly report data (January 2010 to November 2015) from two companies. The study's methodology involves extracting 67 features from a developers' first six months of monthly reports, spanning six dimensions. The researchers used multiple classifiers such as Naive Bayes, Support Vector Machine, Decision Tree, K-Nearest Neighbor, and Random Forest. These classifiers were trained using the Weka tool and 10-fold cross validation. The researchers were able to evaluate the performance of the classifiers using accuracy, precision, recall, F1-score, and AUC metrics. Ultimately the prediction models were compared against a random prediction baseline model which had an accuracy, AUC and recall score of 0.50.

The results from this paper show us that the Random Forest classifier had the best performance with the highest accuracy, F1-score, and AUC among the various classifiers used. The Random Forest model was able to achieve over 70% accuracy on all datasets. The study also shows that the top three most important factors influencing the random forest model include the mean number of tokens in a task report, the standard deviation of working hours, and the standard deviation of working hours of project members in the first month. This tells us that developers who provide more detailed task reports, have less stable working hours, and experience an imbalance of working hours among project members in the first month are more likely to leave the company within their first year of employment. These results highlight the significant impact of the working environment on developer turnover.

In our research we sought to understand the factors that can influence the salary of a software developer and attitude towards blockchain technology using survey data from about 70,000 software developers. Our objective was to develop predictive models, with two focusing on salary determinants and two focusing on blockchain technology favorability. Our approach to constructing predictive models bears some similarities. In our study we developed two linear regression models for salary determinants, as well as logistic regression and Naive Bayes models for blockchain favorability using a wide range of factors.

Proposed Models:

In this section we will propose 4 models. The two first models utilize linear regression to give insights into factors affecting developer's salary. The last two models use logistic regression and Naive Bayes to predict whether a developer would be favorable toward Blockchain technology. All of these models are trained with a 10-fold repeated Cross Validation process using the Caret package in R.

1. First Model:

For the first model, we tried to model developers' salary using linear regression. Our proposed equation can be written as follow:

$$\begin{aligned} ConvertedCompYearly &= \beta_0 + \beta_1 USA + \beta_2 YearsCodePro + B_3 India + \beta_4 VerySmallComp + \beta_5 FullInPerson \\ &+ \beta_6 MasterEdu + \beta_7 Britain + \beta_8 Brazil + \beta_9 NumOfDevType + \beta_{10} VeryLargeComp + \beta_{11} Canada + \beta_{12} NumOfLanguages \\ &+ \beta_{13} Germany + \beta_{14} College + \beta_{15} SmallComp + \beta_{16} LargeComp + \beta_{17} Hybrid + \beta_{18} DataScientist + \beta_{19} Executive \end{aligned}$$

Here are our independent variables and our expectations of their relationship with the dependent variable:

- **USA:** This is a dummy variable that is true if the developer is in the USA. We suspect that as the United States pays developers much higher than other countries, this variable will positively impact their salary.
- **YearsCodePro:** This variable determines the number of years the developer has been coding professionally. We expect that the longer a developer has been coding professionally, the higher salary they have.

- **India:** This is a dummy variable that is true if the developer is in India. We suspect that as India is very advanced in software technologies, the developers will be paid higher than the ones not in India, keeping other variables constant.
- **VerySmallComp:** This is a dummy variable that is true if the developer is working in a very small company, as we have defined in the Data section. We expect small/start-up companies will not have a big budget, so the developer will be paid less.
- **FullInPerson:** This is a dummy variable that is true if the developer is working fully in person. We expect that a developer will be paid more if he is working in person.
- **MasterEdu:** This is a dummy variable that is true if the developer has a Master's degree. We expect that having a Master's degree will improve a developer's salary.
- **Britain:** This is a dummy variable to indicate if the developer is in Britain or not. We expect that this factor will positively affect the salary of a developer.
- **Brazil:** This is a dummy variable to indicate if the developer is in Brazil or not. We expect that this factor will negatively affect the salary of a developer.
- **NumOfDevType:** This is a variable showing the number of titles of developers that the developer has. We expect it to have a positive effect on the salary of the developer.
- **VeryLargeComp:** This is a dummy variable that is true if the developer is working in a very large company, as we have defined in the Data section. We expect large companies will pay developers higher than their counterparts.
- **Canada:** This is a dummy variable indicating if the developer is in Canada or not. We expect it to have a positive impact on the salary of the developer.

- **NumOfLanguages:** This variable shows the number of programming languages that the developer has worked with. We expect that the more languages a developer knows, the more competitive he is and thus the more salary he gets.
- **Germany:** This is a dummy variable indicating if the developer is in Germany or not. We expect it to have a positive impact on the salary of the developer.
- **College:** This is a dummy variable indicating if the developer has gone through College or not. We expect it to positively influence the salary.
- **SmallComp:** This is a dummy variable that is true if the developer is working in a very small company, as we have defined in the Data section. We expect small companies will not have a big budget, so their developers will be paid less.
- **LargeComp:** This is a dummy variable that is true if the developer is working in a very large company, as we have defined in the Data section. We expect large companies will pay developers higher than their counterparts.
- **Hybrid:** This is a dummy variable indicating if the developer works hybrid or not. We expect it to have a positive impact on the salary.
- **Data Scientist:** This is a dummy variable indicating if the developer is a data scientist or not. We expect it to have a positive impact on their salary.
- **Executive:** This is a dummy variable indicating if the developer is in a Senior executive position (Managers, C-Suites, etc). We expect it to have a positive impact on the developer's salary.

2. Second Model:

The second model takes the same inspiration from the first model but with a small modification. The equation can be expressed as:

$$ConvertedCompYearly = \beta_0 + \beta_1 USA + \beta_2 YearsCodePro + \beta_3 YearsSq \ + B_4 India + \beta_5 VerySmallComp + \beta_6 FullInPerson$$
$$+ \beta_7 MasterEdu + \beta_8 Britain + \beta_9 Brazil + \beta_{10} NumOfDevType + \beta_{11} VeryLargeComp + \beta_{12} Canada + \beta_{13} NumOfLanguages$$
$$+ \beta_{14} Germany + \beta_{15} College + \beta_{16} SmallComp + \beta_{17} LargeComp + \beta_{18} Hybrid + \beta_{19} DataScientist + \beta_{20} Executive$$

In this model, we include all of the variables in the first model and add another variable called **YearsSq**, which is the square of the **YearsOfCodePro** variable. We expect that the relationship between YearsOfCodePro and Salary can be plotted as a concave down graph, as later on in their career developers might not need a high salary anymore, and focus more on their life outside of their professional career. The graph may thus go up to local maxima and then come down.

3. Third Model:

Our third model tries to predict the opinion of a developer regarding blockchain technology. The model uses logistic regression. Due to the length of the equation, we will express the equation in the form of $ln(P/(1 - P))$ instead of the form $P = e^{a+bx}/(1 + e^{a+bx})$:

$$ln(BlockchainFavour/(1 - BlockchainFavour)) = \beta_0 + \beta_1 NumOfTech + \beta_2 NumOfWebframe + \beta_3 ConvertedCompYearly$$
$$+ \beta_4 Young + \beta_5 Retiring + \beta_6 Senior + \beta_7 Normal + \beta_8 NewComp + \beta_9 BackendExposure + \beta_{10} SolidityWantTo$$
$$+ \beta_{11} WorkWithSolidity + \beta_{12} Entr + \beta_{13} BlockchainDev$$

To do so, we proposed some of the following variables:

- **NumOfTech:** The number of technologies a developer has used in their career. We believe that as a developer uses more technologies, they will be more likely to open to new technologies such as blockchain, and thus will have a positive impact on their view of blockchain.
- **NumOfWebframe:** The number of web frameworks a developer has used in their career. As web frameworks constantly change, we believe that as a developer is open to learning

different web frameworks, they will be more likely to be open to new technologies such as blockchain, and thus will have a positive impact on their view of blockchain.

- **ConvertedCompYearly:** The salary of the developer, in US dollars. We suspect that as a developer gains more salary, they are more competent and knowledgeable regarding technology in general, and may be able to see the downsides of blockchain, and also not be affected by the hype. Thus it is expected to have a negative impact on the dependent variable.
- **Young:** The dummy variable indicates if the developer is from 18 - 24 years old. We expect that the younger developers are more likely to open to blockchain and catch up with the current hype, and thus having a positive impact on their view of the technology.
- **Normal:** This dummy variable indicates if the developer is in the Normal category, as we defined in the Data section. We expect it to have a positive impact on the blockchain favor, compared to the Middle category.
- **Retiring:** This dummy variable indicates if the developer is in the Retiring category, as we defined in the Data section. We expect it to have a negative impact on the blockchain favor, compared to the Middle category.
- **Senior:** This dummy variable indicates if the developer is in the Senior category, as we defined in the Data section. We expect it to have a negative impact on the blockchain favor, compared to the Middle category.
- **NewComp:** This dummy variable indicates if the developer is working in a company with 2 - 99 employees (a combination of Very Small to Small companies, excluding Freelancers).

- **BackendExposure:** This dummy variable is true if a developer has worked as a backend developer. As developers with backend experience will be more likely to understand the technology, we expected them to not follow the hype and thus may have a negative impact on their blockchain favor.
- **SolidityWantTo:** This dummy variable indicates if a developer wants to work with Solidity in their future career. Solidity is a programming language to develop applications on various blockchain platforms, so we expect people who want to work with it will be more favorable of blockchain, and thus have a positive impact.
- **WorkWithSolidity:** This dummy variable indicates if a developer has worked with or is currently working with Solidity. Similar to the previous dummy variable, we expect it to have a positive impact on the perspective of the developer on blockchain.
- **Entr:** This dummy variable indicates if the developer is bootstrapping a business as a part of their coding activities. We believe that developers with an entrepreneurial mindset will be more open to new technologies like blockchain, and thus have a positive impact on their perspective on this technology.
-

4. Fourth model:

For the fourth model, we also try to predict if a developer would like blockchain or not, using a Naive Bayes model. Let

$S = \{Aspiring,\ BlockchainHobby,\ EntrWithBlockchain,\ Retiring,\ Senior,\ Normal,\ Middle,\ NewComp,\ NewCompWithBlockchain,\ BackendExposure,\ Hobby,\ StrictlyWork,\ SolidityWantTo,\ WorkWithSolidity,\ Entr,\ BlockchainDev\}$

be the set of independent events. Those will be the events that will be used in the Bayes Theorem to build the Naives Bayes Model.

The model has relatively similar independent variables with our third proposed model. We removed the **NumOfTech**, **ConvertedCompYearly** and **NumOfWebFrame** variables as they are continuous and hard to categorize. We then add 6 more variables, which are explained below:

- **Hobby:** This variable is true if the developer does code as a hobby.
- **Aspiring:** This dummy variable is true if the developer is young and wants to work with solidity. In other words, $Aspiring = Young \times SolidityWantTo$
- **BlockchainHobby:** This dummy variable is true if the developer wants to work with Solidity and they code as a hobby.
- **EntrWithBlockchain:** This dummy variable is true if the developer has entrepreneurial mindset and wants to work with blockchain.In other words, $EntrWithBlockchain = Entr \times SolidityWantTo$
- **NewCompWithBlockchain:** This dummy variable is true if the developer works in a new company and they work with Solidity. In other words, $NewCompWithBlockchain = NewComp \times WorkWithSolidity$
- **StrictlyWork:** This dummy variable is true if the developer only does programming in their work, and not outside of work.

Empirical Methodology/Estimation Results:

Using R with caret package to train the models with 10-fold Cross Validation, we got the following results:

	Model 1		Model 2	
	Estimate	P-Value	Estimate	P-Value
Intercept	27751.8	2.00E-16	17910.448	2.00E-16
USA	82311.1	2.00E-16	83223.959	2.00E-16
YearsCodePro	1832	2.00E-16	4548.442	2.00E-16
YearsSq	-	-	-86.486	2.00E-16
India	-20480	2.00E-16	-18788.701	2.00E-16
VerySmallComp	-8909	2.00E-16	-7027.238	1.31E-15
FullInPerson	17686.7	2.00E-16	15072.646	2.00E-16
MasterEdu	7550.2	2.00E-16	6471.554	2.00E-16
Britain	27358.8	2.00E-16	27598.134	2.00E-16
Brazil	-21758.4	2.00E-16	-21796.059	2.00E-16
NumOfDevType	-1298.8	2.00E-16	-1589.776	2.00E-16
VeryLargeComp	12755.6	2.00E-16	13000.173	2.00E-16
Canada	34548.9	2.00E-16	34605.434	2.00E-16
NumOfLanguages	415.3	2.58E-04	364.395	0.00119
Germany	10451.1	2.00E-16	9892.957	2.00E-16
College	1214.3	5.93E-02	753.778	0.23677
SmallComp	-4636.9	8.16E-08	-3936.022	4.20E-06
LargeComp	7522.4	3.58E-13	7523.591	1.96E-13
Hybrid	14251.2	2.00E-16	12404.388	2.00E-16
DataScientist	7439	5.15E-08	8477.576	3.56E-10
Executive	35366.4	2.00E-16	34579.223	2.00E-16
	Model 1		Model 2	
RMSE	56575.68		55928.71	
R-Squared	0.3802687		0.3944581	

Adjusted R-Squared	0.3805	0.3939
MAE	35610.88	34931.93
P-Value	2.20E-16	2.20E-16

Based on the result, we can see that for the two linear regression model, every independent variable has a p-value of less than 0.05, except for the College variable. That means that except for the College variable, every independent variable is statistically significant at a 5% significance level. The two models also have a p-value of 2.2E-16 each, so the variables are jointly statistically significant.

Between the two linear regression models we ran the optimal model to predict the wage of a software developer would be Model #2. We can say this with confidence because Model #2 has a higher R-squared and lower RMSE and MAE scores (we obtained these figures using 10-fold cross validation) which tells us that the model has greater explanatory power over the dependent variable.

For Model #3, we got the following result:

Model #3	Estimate	P-Value
Intercept	-1.302E+00	2.00E-16
NumOfTech	1.201E-02	6.77E-06
NumOfWebFrame	3.017E-02	0.000403
ConvertedCompYearly	-1.448E-08	0.381306
Young	3.421E-01	2.00E-16
Retiring	-1.014E-01	0.03538
Senior	-2.490E-01	0.001664
Normal	1.529E-01	1.05E-07

NewComp	5.860E-02	0.022731
BackendExposure	-5.523E-02	0.020159
SolidityWantTo	2.368E+00	2.00E-16
WorkWithSolidity	6.980E-01	5.60E-06
Entr	4.997E-01	2.00E-16
BlockchainDev	2.315E+00	2.00E-16
Model Accuracy	0.737149	

The third model also have a relatively good set of independent variables. Except for the ConvertedCompYearly variable, all independent variables have a p-value of less than 0.05, which is our significance level.

We can also analyze the confusion matrix of the third model, which is given as follow:

	True value of 0	True value of 1
Predicted value of 0	69.8%	25.5%
Predicted value of 1	0.8%	3.9%

As observed in the confusion matrix, the majority of the failed predictions of the model is Type II error (false-negative), with a probability of 25.5%. This suggests that there are other factors that affect the developer's perspective on blockchain that we have not included in the model. That is also an improvement for the model.

The third model here also has an accuracy of 73.71%, which is acceptable as it is better than a random model (50% accuracy) and it correctly predicts the majority of time. However, improvements should be made to increase the accuracy to 80 – 85%.

For the fourth model, we got an accuracy of 72.56% with the following confusion matrix:

	True value of 0	True value of 1
Predicted value of 0	68.4%	26.5%
Predicted value of 1	0.9%	4.1%

As we can observe, the fourth model has a relatively similar accuracy and confusion matrix with the third model. This suggests that there are factors omitted in the two models that should be considered. The fourth model also performs with an acceptable result, yet improvements can be made to increase the accuracy rate of the model.

Regarding comparisons between Model #3 (the logistic model) and Model #4 (the Naive Bayes model) the optimal choice to predict blockchain favorability would be Model #3 as it has a greater accuracy score, and less error.

Conclusion:

The objective of this study is to provide insights into the factors that impact a software engineer's salary and attitudes towards blockchain technology. The industry for developers has grown tremendously, we were interested in building accurate models to predict these characteristics among developers.

In our most accurate model in predicting software engineer salary, the 2nd model, we found twenty-one variables that had a significant impact on expected salary. This model has an R-squared of .3944, which tells us that we are 39.44% of the way towards being able to perfectly predict a developer's salary. We found locations have a significant factor in the estimated salary, specifically USA, India, Brazil, Canada, and Germany, which all improved our model. Developers in the USA showed the highest salary increase. We also discovered that company size is an important factor in predicting salary, and including this variable in our model significantly improved its accuracy. Specifically, employees in large and very large companies tend to earn higher salaries. Other key factors influencing salary include the employee's job role, with Executive and Data Scientist positions earning more, as well as their work arrangement, such as working fully in-person, in a hybrid setting, or fully remote.

In blockchain favorability for software engineers, we found fourteen coefficients that had a significant impact for our most accurate model, the 3rd model. This model has an accuracy of 73.71% in predicting blockchain favorability. We found significant variables for predicting favorability when looking at the age of the developer, with young developers more favorable to blockchain. Developers that are currently working with and want to work with the technology Solidity, also are favorable for blockchain. Our Entrepreneurial dummy variable also showed favorability towards blockchain technologies, which suggests high blockchain favorability among entrepreneurial developers. Other than that, developers with backend development experience, number of technologies used, and web frameworks all have a significant impact on predicting the favorability.

As a closing note, we believe the insights from our compensation models would be helpful for companies recruiting software engineers and for developers to find ways they may be able to increase their expected salary within the industry. On the other hand, our blockchain favorability models will be beneficial for the adoption of blockchain and help this technology gain a better outlook from the developers.

Works Cited:

Bhat, Dheemanth. “Stack Overflow Annual Developer Survey 2022.” *Kaggle*, 29 July 2022, https://www.kaggle.com/datasets/dheemanthbhat/stack-overflow-annual-developer-survey-2022.

Lingfeng, Bao, et al. “Who Will Leave the Company?: A Large-Scale Industry Study ... - IEEE Xplore.” IEEE Explore, 3 July 2017, https://ieeexplore.ieee.org/abstract/document/7962366.