Recommender System: Finding the True Vertical Depth

Sanya Srivastava, Samira Ravilisetty, Joshua Dierker, Srikanth N

The University of Texas at Austin

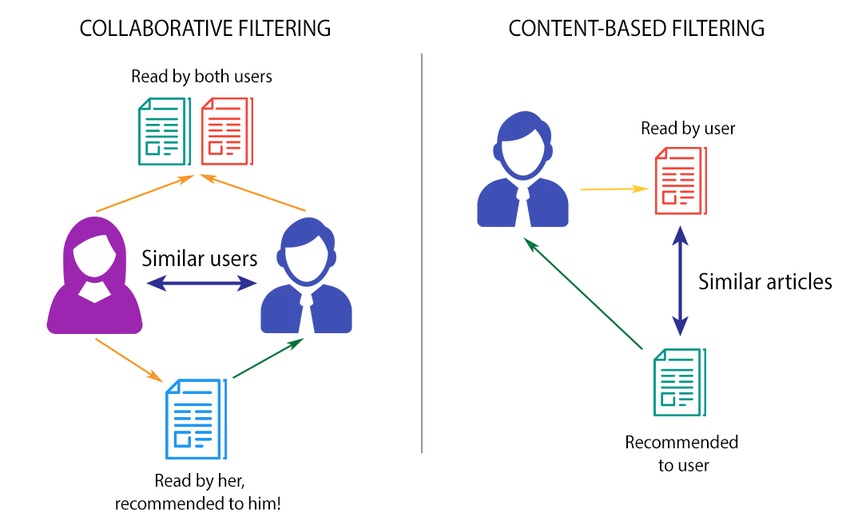
Recommender Systems: Finding the True Vertical Depth

# Abstract

Recommender Systems make suggestions based on various matrix factorization (MF) algorithms, such as Alternating Least Squares (ALS) or Stochastic Gradient Descent (SGD). (Xie, 2019) This paper explores the goal of using recommender systems based on ALS and collaborative filtering (CF) to recommend the True Vertical Depth (TVD) when drilling certain oil well formations using factors such as depth and formation. After standardizing the values for these two factors, we proceeded to implement the matrix factorization using ALS and collaborative filtering - first with the Wyoming dataset and then with the Norway dataset. The Norway model was able to use these machine learning techniques to get a low error of 23 feet between the actual and predicted depth.

Introduction

Recommender systems analyze the relationship between a user and a product and attempt to make accurate predictions based on the data collected from these interactions. Many companies, such as Spotify and Netflix, utilize recommender systems to make recommendations that will increase customer retention and maximize profits. Collaborative filtering is for relatively larger datasets, such as the Wyoming and Norway oil well data (Luo, 2018). Unlike content-based methods, collaborative filtering focuses more on “similarity from [user] interactions” rather than the similarities between items (Kordik, 2018). The main purpose of collaborative filtering is to narrow down a user’s interests depending on the reactions of users similar to them and then make predictions based on what the user might be interested in. We will be using the latent factor method instead of the neighborhood method for our datasets. The neighborhood factor method focuses more on comparing the relationship between items and users and makes predictions based on the ratings of similar objects by a user (Luo, 2018). On the other hand, the latent factor method focuses on the factors that the user has an interest within an item. This paper focuses on the latent factor method because the latent factor method does not look at the user as a whole since excessive information may be prone to error (Xie, 2019). Focusing on a particular factor allows dimensions to reduce for more accurate predictions. This paper focuses on the challenge of utilizing the methods mentioned above to implement matrix factorization to create a recommender system that will predict the True Vertical Depth (TVD) for future oil well formations.



**Figure 1 (Luo, 2018)**

**Comparison between Collaborative and Content-Based Filtering**

Background

The True Vertical Depth (TVD), or the vertical distance from the ground level of the well, is an essential factor that engineers have to consider when drilling into a specific formation. TVD can be challenging to measure because the borehole could be contorted, and the points that are being analyzed within the formation are placed close to each other. Using the provided Wyoming and Norway dataset, we hope to make accurate predictions of the TVD based on various values of depths and formations given to us. With ALS, we will be able to optimize oil and gas production.

In this dataset, we have Wyoming, which is a dataset that has the most challenging structure. We start by standardizing the formation by changing the formation into numerical values using a mapping function in python. Then, we use the preprocessing library to standardize formation. After that, we work on standardization of depth, which was done by taking measured depth (MD) and subtracting the kelly-bushing, an elevated device on top of the ground-level. We are able to loop through the depths provided to create a new TVD column. This same standardization model was applied to the Norway model as well. Finally, we are given a top depth recommender system model that we had to apply to Wyoming and Norway data sets. We create a pivot table of the well identification (API number or name) and the standardized formation. Within that table was a standardized depth. After the creation of the pivot table, we normalize the data that we have and run it through some iterations in the ALS model.

The data that is associated with recommender systems are based on both explicit and implicit feedback. Explicit feedback is when a consumer expresses their interests through activities, such as rating a product out of five stars (Luo, 2018). On the other hand, implicit feedback is when a consumer shows their opinion by activities such as viewing, searching for, or purchasing an item. The data for the recommender system is in the form of a matrix which holds the reactions of the user for each item. The matrix can be either sparse (one with empty cells) or dense (mostly filled) (Ajitsaria, 2019). Alternating Least Squares (ALS) is a matrix factorization algorithm that is efficient and useful for a sparse matrix, like the matrices that we deal with for Wyoming and Norway. ALS minimizes two different loss functions; it holds the “user matrix fixed and runs gradient descent with item matrix; then it holds item matrix fixed and runs gradient descent with user matrix,” (Liao, 2018). As a result, ALS helps create a model that is reliable with low error. We measure model error by mean absolute error (MAE), which is used from the sklearn library. MAE is the average magnitude of errors without the consideration of errors. This error method will be used in the project because it has a focus on the overall data average and can have an easy interpretation. (JJ, 2016).

# Methods

Cleaning/Preprocessing

We standardize the true depth data and formation from the Wyoming Data set. This standardization can allow our data to be consistent with the content and format of our results. After the standardization of total depth, we incorporate the data with the formation of the wells. Since formation is a qualitative data type, we convert each measured data to an assigned numerical value. Then, we standardize the formation of that data for consistency (Luo, 2018). We combined the names of specific data with its corresponding formation and total depths to create a new data set for matrix factorization.

True Vertical Depth, for all wells, in our data set found by finding the depth of each well from a standard point of sea level. In the dataset for the Wyoming database, all wells are situated above sea level. There were three points of measurement from which Total Depth calculated: Kelly-Bushing height, Drilling Floor height, and Ground Level height. Finding True Vertical Depth involves standardizing measurements of depth from sea level, so we found this through the formula:

True Vertical Depth = Height of Measured Point - Total Depth from Measured Point

A TVD measurement is found by subtracting the ground elevation from the total depth if ground elevation is positive, and adding ground elevation to total depth if ground elevation is negative. This virtually ensures that all measurements of depth are standardized from a 0-foot measure, and ensures that TVD is accurate unless the well is not conventional (one of our assumptions is that the well is standard).

Collaborative Filtering with ALS

Collaborative Filtering is typically used in large data sets. In this case, our Wyoming data was large and had many corresponding factors that can be used in matrix factorization.(Luo, 2018) The total depth and formation were the two factors in the matrix factorization. The total depth of the formation tops () is the result we wanted to attain from understanding the formation of the well and was used in the equation:

where is the total depth as that is the item of interest, is the formation that has an “interest” in the total depth, is the recommended TVD that is the product of the matrix factorization (Rackaitis, 2019). We are using ALS to conduct the factorization. This algorithm works by minimizing the loss function of the form:

The loss function, L, is an error metric of how well the algorithm fits the data (Rosenthal, 2016). The minimization of the loss function was important because the smaller the loss, the lower the error in our recommendation. This algorithm does a gradient descent first using the total depth (TD) matrix while holding the formation matrix constant and minimizing the loss function.(Rackaitis, 2019) We decide to use the ALS algorithm because it performs well with large amounts of data and is able to handle cases where the data is sparse such as ours, where we had several NaN entries that were cleaned from the data set. This would ideally reduce the error in our recommendations (Rosenthal, 2016). However, the values for the formation were in the form of an array, so we convert it into its own separate data frame. From there, we made sure that the values in both of the data frames were aligned by the API Number of each oil well, to ensure that the data frames match each other.

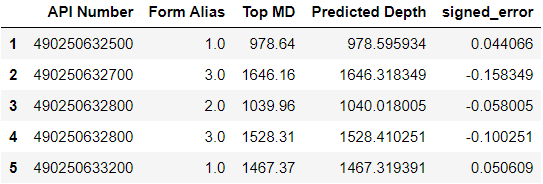
Data Visualization and Error

Finally, we have the predicted and actual data frame that needs to be visualized then converted back into the testing data frame to apply MAE. We use matplotlib to visualize the graph, with a 1:1 line. After that visualization, we reform the pivot-table to look like the test data frame we had before-hand. This involved using pandas and stack methods to reshape the data frame. Once the final data frame is complete, we apply the MAE error from the sklearn library to take the actual TVD and predicted TVD.

# Results

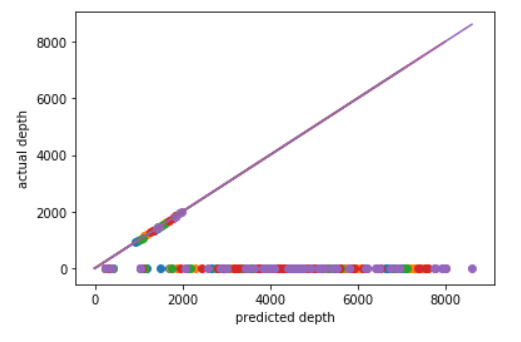
Wyoming Data Set

The predictions of the Wyoming model result in low error. At the end of 20 iterations of our factorization model, the error in solving for the user matrix is 0.727 and the error in solving for our item matrix is 0.409. The figure below showcases our actual formation top depths as “Top MD” and our predicted formation top depths as “Predicted Depth”.



**Figure 2: Head of DataFrame showcasing actual and predicted depths**

Our predictions are very close to the actual formation top depths, with the majority of the predictions landing within hundredths of a foot of the actual. In the graph below, our predictions lie almost perfectly on the 1:1 line of actual vs predicted values, meaning that the predictions are almost exactly the actual formation top depths.

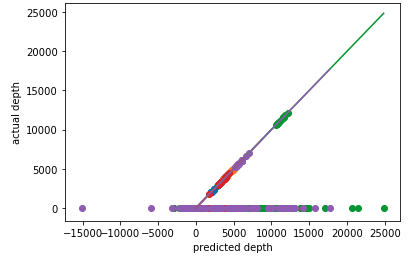


**Figure 3: Plot of Actual formation top depth vs predicted depths with 1:1 line**

The overall mean absolute error in our model is 1.015. The error that is inherent in the model lies in disparities in the subsurface. For example, we may be predicting depths for formation tops that do not span to the Northing and Easting we are predicting at, leading to extrapolated results that are ultimately meaningless.

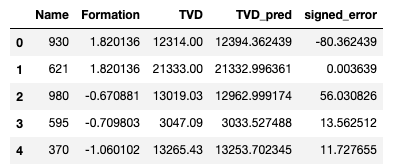
Norway Data Set

The application of the Norway model involves similar steps as Wyoming. After cleaning and standardizing the data, we move on to implementing the matrix factorization. Using ALS, the error seems high on the Norway data, however, after each iteration, the error continues to become smaller and smaller. We then proceed and transform the standardized values back to the original values and plot the graph onto predicted depth versus actual depth, as shown below.



**Figure 4: Graph of Predicted Depth versus Actual Depth Line Space**

Based on this graph, we are able to see that the predicted depth in the data was identical to the actual depth since they fall on the 1:1 line, which is similar to a statistical line of best fit. This makes the predictions more aligned to be accurate. This new information on the graph then transitions into a data frame.



**Figure 5: First 5 wells of Actual TVD versus Predicted TVD**

The data frame above includes a “signed\_error” column which signifies the difference between actual and predicted values. We then apply the mean-absolute-error to find how far apart the target values are for actual and predicted, which utilizes the error column. We find that the Norway model had a mean-absolute-error of 23 feet between actual and predicted.

# Discussion

As shown in the figures above, the results that we received had a low error. Each of the graphs displayed have a strong positive correlation between the predicted depth and the actual depth which signifies that the values are close to each other. However, this model is only applicable for datasets in which all the data is known; it is unable to predict the accurate TVD for wells that have unknown or empty values. Despite this limitation, considering that a substantial number of wells had unknown values, our matrix factorization predicts depths with extremely low error values that decreased with each iteration for both the Wyoming and Norway data sets. Additionally, the model for the Wyoming data set consistently overpredicts depths by a small margin, while the model for the Norway data set has a more even distribution of over and under predictions. Overall, the results aforementioned indicate that utilizing matrix-factorization based ALS returns significantly accurate values, due to the fact that all the error values are relatively low and that the predicted versus actual depth have a strong correlation when we plot them against each other. Our model is a trustworthy method that can be used to predict the TVD when drilling oil wells in the future.

Conclusion

Matrix-factorization based recommender systems are useful in many industries to make item recommendations to users to ensure consumer satisfaction. Recommender systems are especially valuable for energy analytics to predict the TVD for future oil wells to maximize oil and gas production. In addition to ALS, we chose to use collaborative filtering, which is helpful when analyzing a large dataset. We then standardized the data, implemented the matrix factorization and ALS, and plotted the values we predicted versus the actual values to compare the two. Our model returned extremely accurate TVD values with low error (Wyoming had 1.015 and Norway had 23.5), which displays through various figures in which we compare the predicted depths to the actual depths. However, more research has to be done to adapt the model to deal with oil well formations with unknown data points and outliers, since that is a common occurrence when dealing with oil wells. Recommender Systems that we have created are valuable in the oil/gas industry; however, it can apply to other sectors. Recommender Systems can be used in the educational field, where a student’s test score on a particular subject causes a recommendation on what textbook to get for them to better their scores. Our recommender system uses the ability to look at large data sets like Wyoming and Norway to predict TVD to optimize oil production. Since our error was low on both models, these recommender systems can apply to many ideas to optimize and profit for a company.

References

Ajitsaria, Abhinav. (2019, July 10). *Build a Recommendation Engine with Collaborative Filtering.*

<https://realpython.com/build-recommendation-engine-collaborative-filtering/>

Jj. (2016, March 23). *MAE and RMSE - Which Metric is Better?*

<https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d>

Kordík, Pavel. (2018, December 15). *Machine Learning for Recommender Systems - Party 1 (Algorithms, Evaluation and Cold Start).*

<https://medium.com/recombee-blog/machine-learning-for-recommender-systems-part-1-algorithms-evaluation-and-cold-start-6f696683d0ed>

Liao, Kevin. (2018, November 17). *Prototyping a Recommender System Step by Step Part 2: Alternating Least Square (ALS) Matrix Factorization in Collaborative Filtering.*

<https://towardsdatascience.com/prototyping-a-recommender-system-step-by-step-part-2-alternating-least-square-als-matrix-4a76c58714a1>

Luo, S. (2018, December 10). *Introduction to Recommender System Approaches of Collaborative Filtering: Nearest Neighborhood and Matrix Factorization*. https://towardsdatascience.com/intro-to-recommender-system-collaborative-filtering-64a238194a26

Rosenthal, E. (2016, March 16). *Explicit Matrix Factorization: ALS, SGD, and All That Jazz*.<https://blog.insightdatascience.com/explicit-matrix-factorization-als-sgd-and-all-that-jazz-b00e4d9b21ea>

Xie, Medford. (2019, March 25). *Neighborhood vs Latent Factors Methods in Collaborative Filter Recommender Systems - Part 1.* <https://medium.com/@medfordxie/neighborhood-vs-latent-factors-methods-in-collaborative-filter-recommender-systems-part-1-9f969c4990b0>