Shaswat Rajput
PSY - 6220-01

Online Auction Analysis Using Bayesian Inference Techniques

ABSTRACT

Auction analysis have gained traction and popularity in research community in past few years. There are many factors that influence auction's performance like bidders per lot, reserve price of item etc. In this paper I explored different variables to address questions like does forward auctions are better than sealed bid auctions, does increasing the number of lots offered has positive or negative impact on price, does the number of bidders per lot has positive or negative effect on Return on Reserve (ROR) price. The data collected for this project is from google dataset archives. The data consists of 10 feature columns namely auctionID, Return on Reserve (RoR) price, proportion of lots (by reserve value) that were successfully sold in the auction event, number of lots offered in the auction, number of different product types offered in the auction, average reserve price over all lots in the multi-lot online auction, the average starting-bid, average number of bidders per-lot, the auction mechanism used and state where auction happened. In this project I aim to fit 4 different linear regression models with weakly regularizing priors, normal likelihood, model with student-t likelihood distribution and hierarchical model using Bayesian approach. After modelling process, I will use WAIC score to compare and find the most robust model. After model comparison, I plan on using the robust model to help answer the

questions mentioned above like using the posterior predictive distribution to quantify the impact of bidder per lot etc.

INTRODUCTION

An auction is usually a process of buying and selling goods or services by offering them up for bid, taking bids, and then selling the item to the highest bidder or buying the item from the lowest bidder. There are mainly 3 type of type of Auction: English Auction, Dutch Auction and Sealed Bid.

1. English Auction [2]:
    A. Forward : It starts with a low starting price and then it proceeds with increments in price submitted by the people who are taking part in auction as buyer. In this auction whoever offer the highest bid wins the auction.
    B. Reverse: It starts with a high price point and proceeds with decrement in price submitted by those who take part in auction. In this auction whoever offer the least bid wins. It usually go on for couple of hours as compared to forward auctions which can go on for days.
2. Dutch Auction [2]:
    A. Forward : It starts with a high starting price and then proceeds with price decrement at particular intervals to a minimum price. Supplier is free to choose choose at any bid during decreasing intervals. For example: Example: You are interested in selling off older machinery, you start will a higher selling price that could make your resale profitable, to a break-even price or to the lowest price you are willing to sell for. The bidders will watch

the decrements and decide whether they want to purchase at that particular price. The first bidder who takes the offer price will win the bid.

B. Reverse: It starts with a low starting price and then it proceeds with increment in price at particular intervals to a higher or maximum set price. Supplier is free to choose at any bid during increasing intervals. For example: You are looking for a Heating, Ventilation and Air Conditioning (HVAC) unit and installation at your 300 sq meter warehouse, you start with the lowest fair price that you are willing to pay. The bidders will watch the increments and decide based on their expertise and pricing, whether they can provide the service for that particular price. The first bidder who takes the offer price will win the bid.

3. Sealed-Bid Auctions:

   A sealed bid is type of auction[1] in which all bidders submit sealed bids to the auctioneer so that no other bidder knows what price they have have bid on. It is an auction where :

   • The bids are sealed, often physically in an envelope, and are all opened at once.

   • Sealed-bid auctions are generally used in bidding for government contracts.

   • Unlike an open bid, where buyers can make multiple bids and compete against each other actively, in a sealed-bid auction, they only get once chance.

DATA DESCRIPTION

The data is collected from google open dataset archives and I hosted it on Kaggle. Since this data doesn't have a verified publisher so I'm not sure about the integrity of the data. This data contains the following feature variables -

1. auction_*id: It's a* unique identifier for a single multi-lot online auction event.

2. RoR: It's the Return on Reserve over all lots sold in the auction event.
3. STR: It's the proportion of lots (by reserve value) that were successfully sold in the auction event.
4. lots: It's the number of lots offered in the auction.
5. avg_*reserve:* It's the average reserve price over all lots in the multi-lot online auction.
6. avg_start_*bid:* It's the the average starting-bid (expressed as a fraction of the reserve bid).
7. BPL: It's the average number of bidders per-lot.
8. auction_mech: The auction mechanism used for the auction event (English Forward, Sealed Bid or Fixed Price).
9. state: state where auction happened.

DATA ANALYSIS AND CLEANING

During the initial analysis of data I found that there were many outliers and few missing values. Based on my knowledge and what I know about the data I tried to remove the outliers in the dataset. First I found out that there were some rows where auction price was in millions which I think were outliers, so I removed the top quantile of the average reserve price. After this I found out that there was a ambiguity in RoR values so I filtered out the top percentile values of RoR.

We know that FIXED PRICE items means items which can be bought directly at the price set by the auctioneer and these items are sold on first come first serve basis. Because of this reason these the RoR on such items is 1, so I filtered these too from the dataset as they will only add noise to the model. I also filtered out those rows where no item was sold in the auction.

After doing the above mentioned cleaning in the dataset, I plotted the distribution plot of the target feature which is RoR.

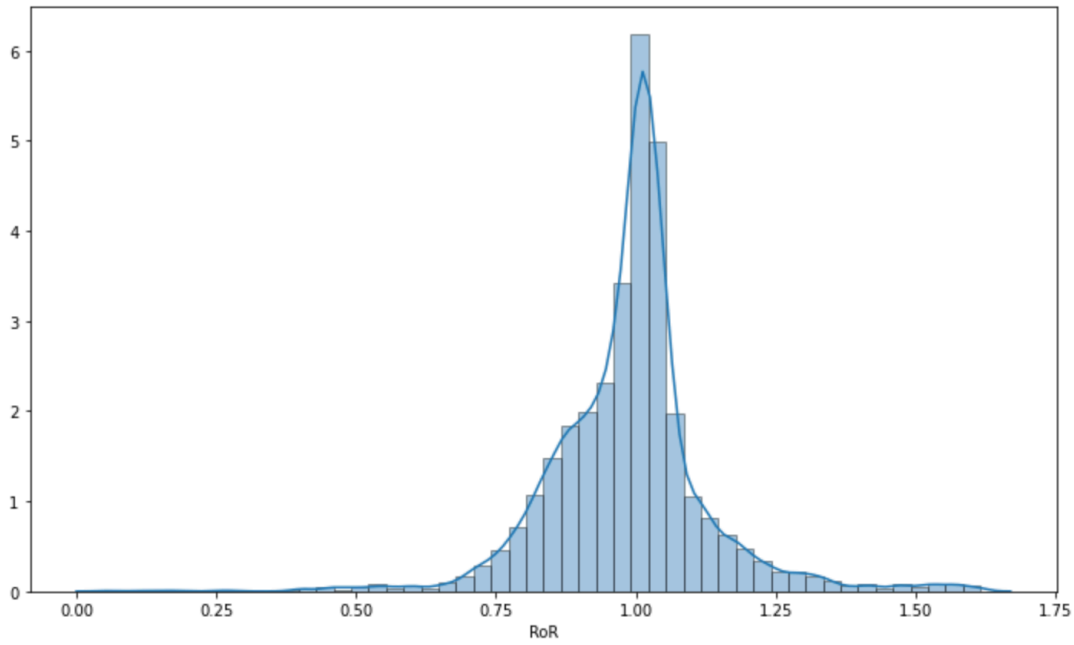

Fig1. Distribution plot of RoR

As we can see in the above plot that RoR seems to have a bi-modal distribution. Since the auction mechanism can have an impact on the RoR, so let's plot the distribution of RoR by the auction mechanism to see if this bi-modality is present in both or not.

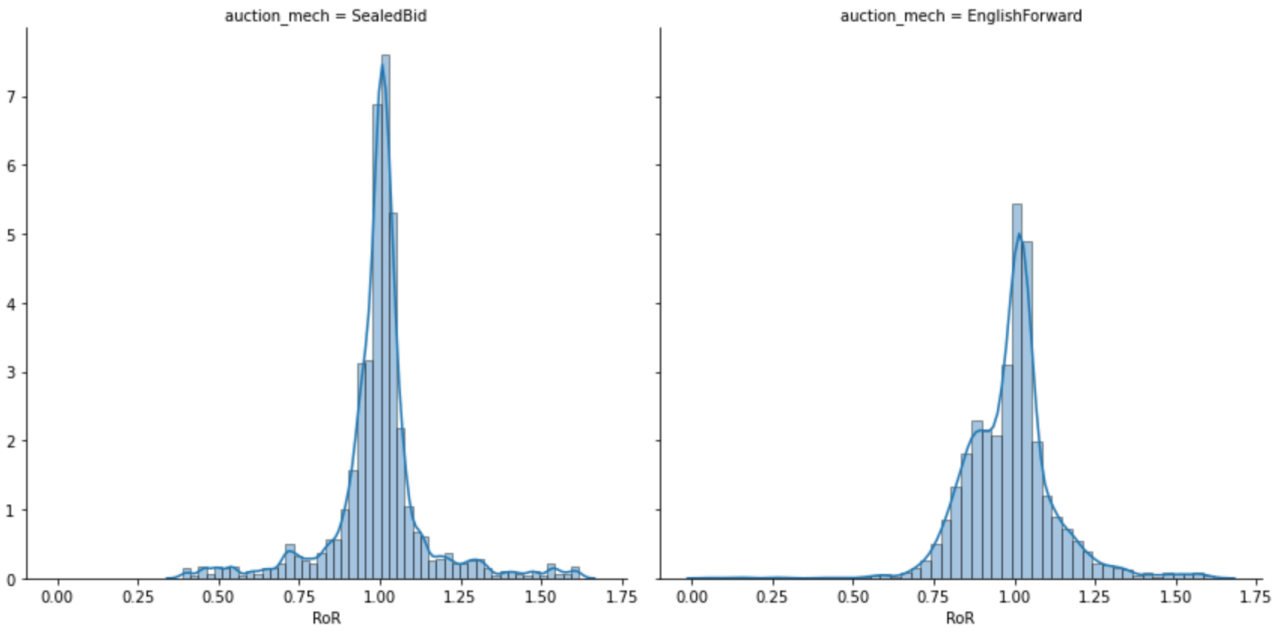

Fig 2. Distribution of RoR by auction mechanism

As we can see in the above plots that the SealedBid seems to have long tails and in this case Student t-distribution might be a good likelihood distribution to use during the modelling phase. Also we can see that English Forward seems to have a bi-model distribution.

Let's explore more about English Forward auction to see which feature might have caused this bi-modality. Since avergae start bid is directly relate to RoR as RoR is nothing but auction price divided by reserve price. Let's plot a scatter plot to see the relation between RoR and average start bid based on states preset in the dataset.

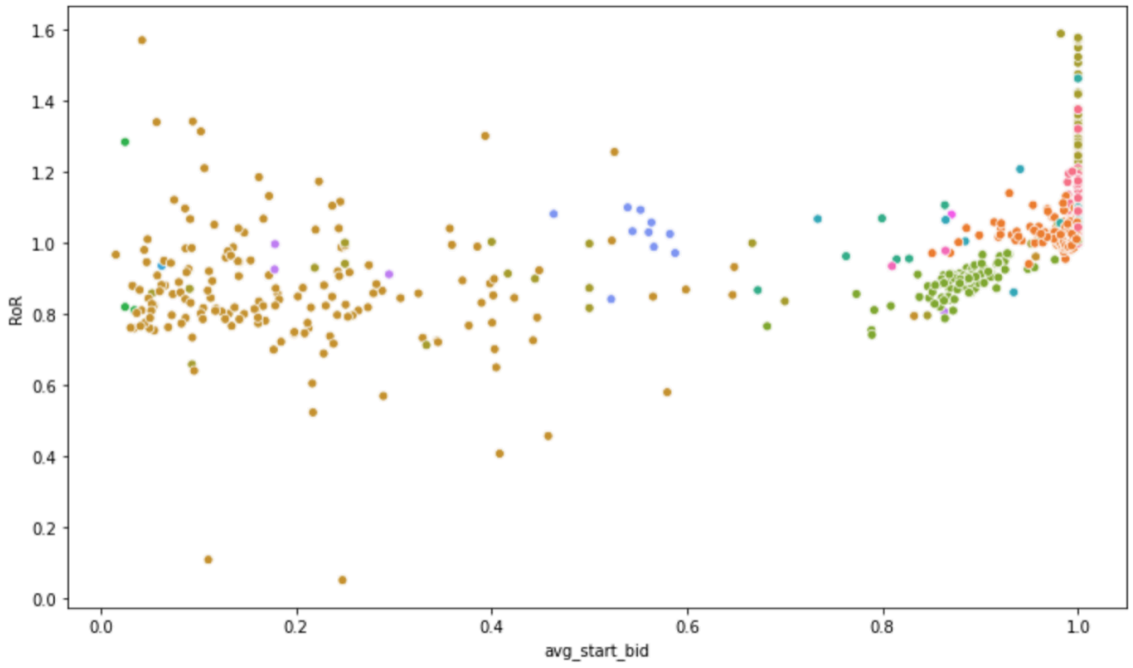

Fig 3. RoR vs average start bid by state

As we can see in above plot that there are some states where RoR is greater than average start bid price which cannot be true so we will filter out these auction events from our data before we move to modelling phase.

The last thing that I wanted to check in our data is if there's any correlation between any feature variable in the dataset.

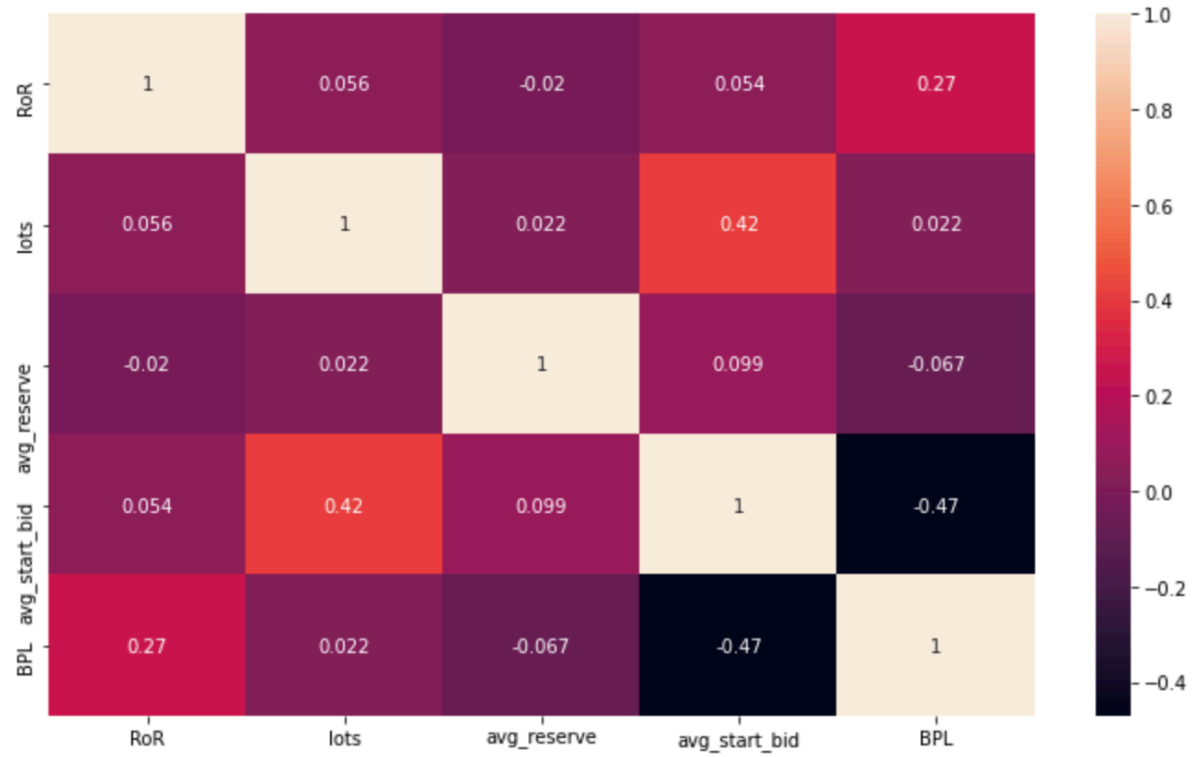

Fig 4. Correlation Heatmap of all feature variables

We can see that from the above heatmap that there is a little positive correlation between average start bid and lots offered in an auction. There is a a little negative correlation between average start bid and bidders per lot feature. So before we start modelling on this dataset, we should take care of this correlation which I will explain in the data pipeline section.

FEATURE SELECTION

Our first goal is to be able to predict well on RoR before we move on to the analysis of variable that affect elections. For this I have chosen the following variables for modelling part -

- Lots - to control the available features.
- avg_reserve - For controlling lot value - e.g. lower value lots may see more bidders willing to go above the reserve and vice-versa
- avg_start_bid - For controlling the public price
- BPL - we know from basic auction theory, that the more bidders there are, then the better expected prices will be.
- auction_mech - whether the lots were traded via English Forward (EF) or Sealed Bid (SB) auctions.
- state - to control for differences between local markets.

DATA PIPELINE

Before modelling, data is needed to be preprocessed. The data pipeline is as following:

- Firstly I selected the variables of interest.
- Then I centred and scale the continuous variables in the dataset.
- Converted the categorical features into Encoding factors.
- Used log transformation on the target feature (RoR).
- And lastly splitting the dataset into training and testing set.

The reason why we centred and scaled the data is because during our data analysis we found the correlation between some features and also by doing centring we are isolating the impact that one input variable can have on the target feature. Also centring and scaling the data by standard deviation makes it easy for interpretation and comparison of parameter estimates.

The reason why we use log transformation on the target feature is because I want to use maximum entropy distribution as likelihood function and by doing the transformation on RoR we then can use maximum entropy functions as they are defined in the range of -inf to +inf. Also by doing log transformation we will able to correctly express the uncertainty in target feature with confidence.

For the last stop in the data pipeline, I used the 75% of the dataset for training purpose and rest 25% of the dataset for testing purpose. I did this step by using the sklearn library's train_test_split method.

MODELING

MODEL 1:

This model is a simple pragmatic linear model in which I -

- Used a Normal distribution for priors and likelihood.
- Used mean 0 and standard deviation 1 for the priors.
- Used Half Cauchy with beta parameter to be 2.

In this model[6][8] I have Normal priors to reflect the maximum uncertainty under the constraint of having a fixed mean and standard deviation. Also a non-flat weakly regularising prior for sigma is chosen for efficient estimation.

The graph model of model 1 is as follows -

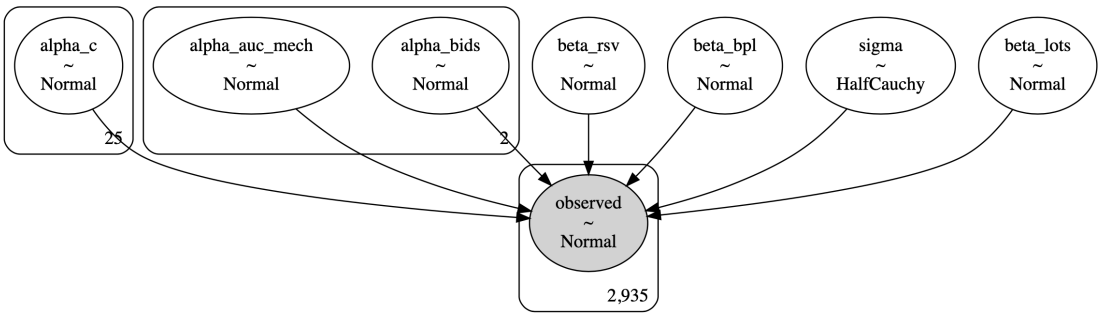

Fig 5. Graph Model of model 1

When printing out the model summary we found that r_hat values in all parameters was 1, mean effective sample size was large and also in plotting the trace we found that there was no divergence. These all prove that our model is acceptable. When I plotted the forest plot I found that spread is large for the variables which is not a good sign. (Trace plot and forest plot are showed in Jupiter notebook).

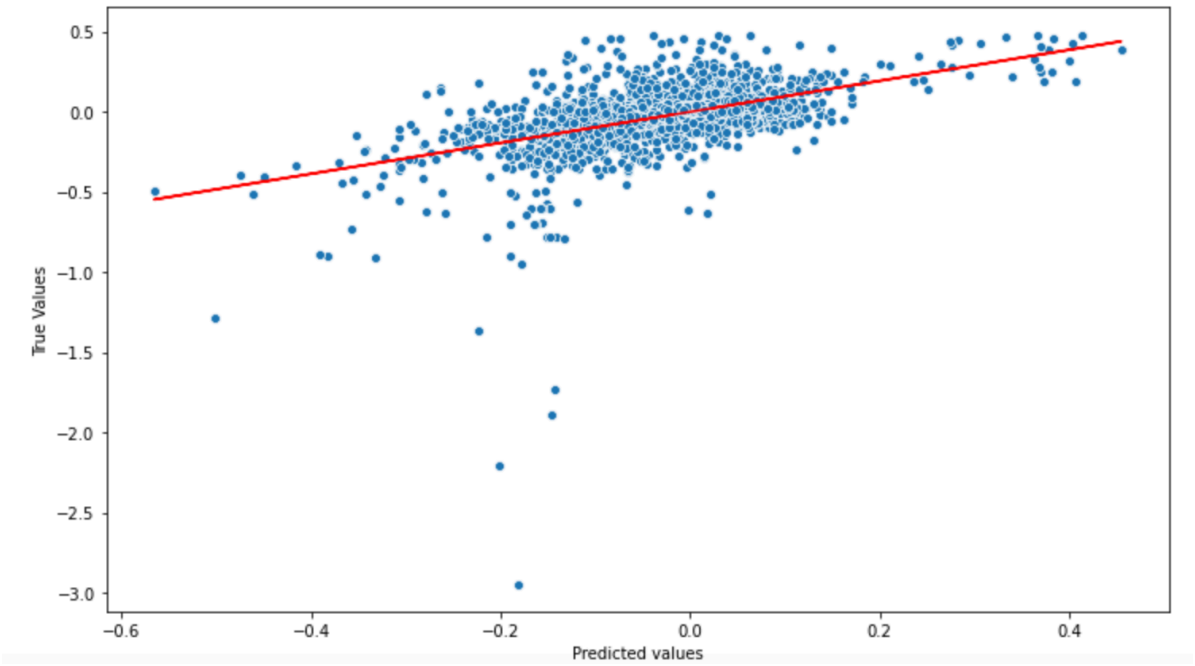

Fig 6. Model Prediction vs True values

As we can see that our model performed ok and that there were some bad predictions too. Let's now plot the residual error distribution to see if your likelihood function choice was ok and that there is no multi-modality.

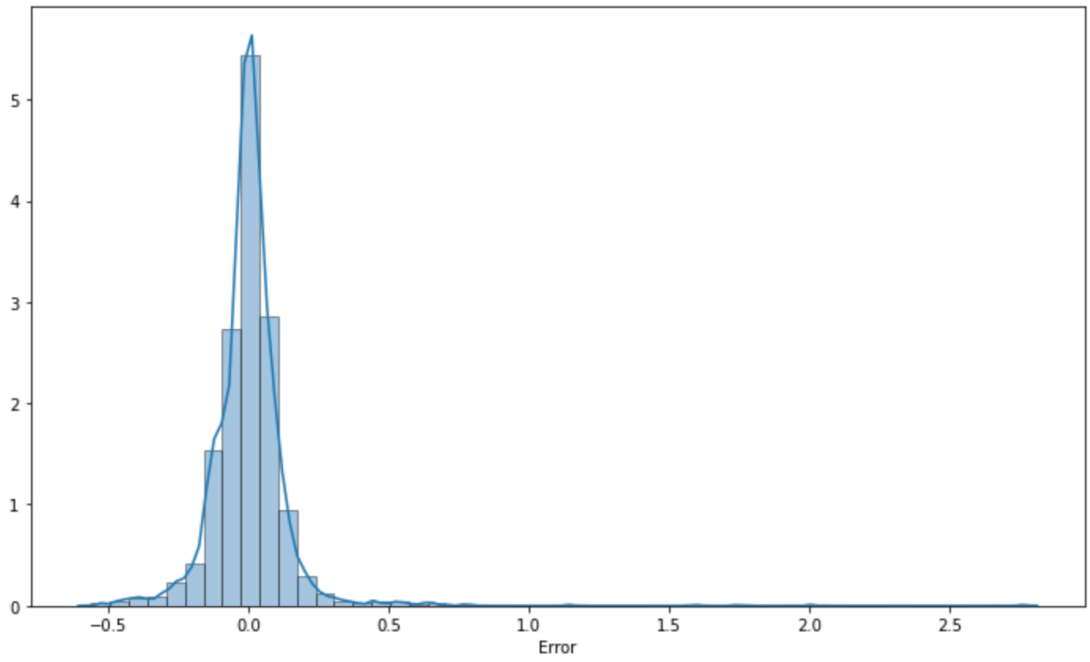

Fig 7. Residual error distribution

It looks like student t-distribution would be a better choice , but before moving on to student t-distribution, let's make some changes to model 1 parameters to see if if it can perform better.

MODEL 2

This model is same as model 1. The only thing that have changed in this model is the value of standard deviation on the prior. I changed it from 1 to 0.2 because I want to have a low probability that a change of 1 standard deviation in independent variable will have impact on the target feature (RoR) . The graph model of this model will be same as of model 1.

When printing out the model summary we found that r_hat values in all parameters was 1, mean effective sample size was large and also in plotting the trace we found that there was no divergence. These all prove that our model is acceptable. When I plotted the forest plot I found that spread is small as compared to model 1 which is also good sign. Let's plot the prediction plot to see if model performed well or not.

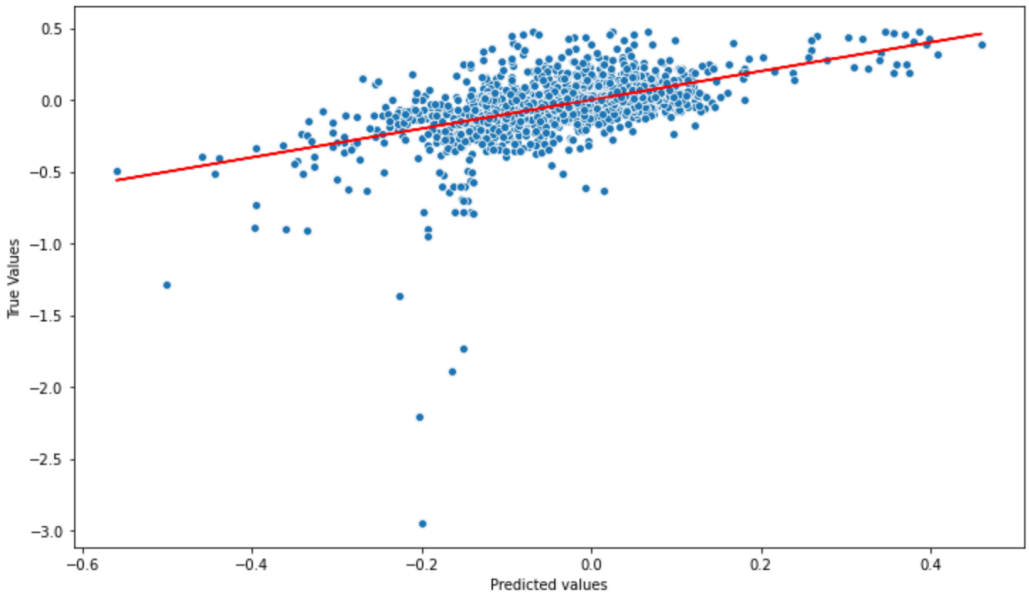

Fig 8. Model Prediction vs True values

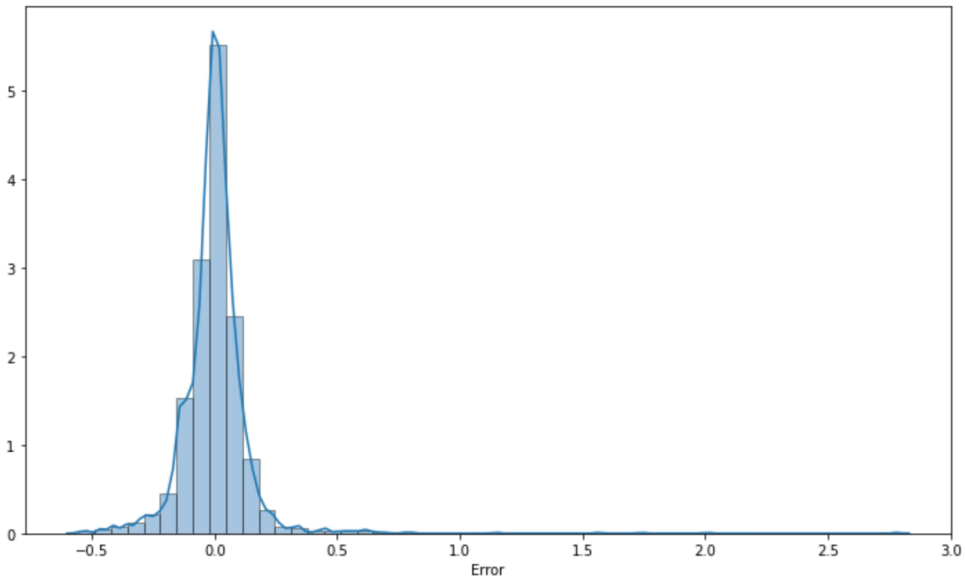

Fig 9. Residual error distribution

As we can see in the above plot that model prediction was ok and that there still were some bad predictions. The residual error distribution shows student t-distribution would a good choice for likelihood distribution to better approximate the fat tails.

MODEL 3

The model parameters in this model is same as of model 2 but the only thing that has changed is that instead of Normal likelihood function we are using student t-distribution[6][9][8].

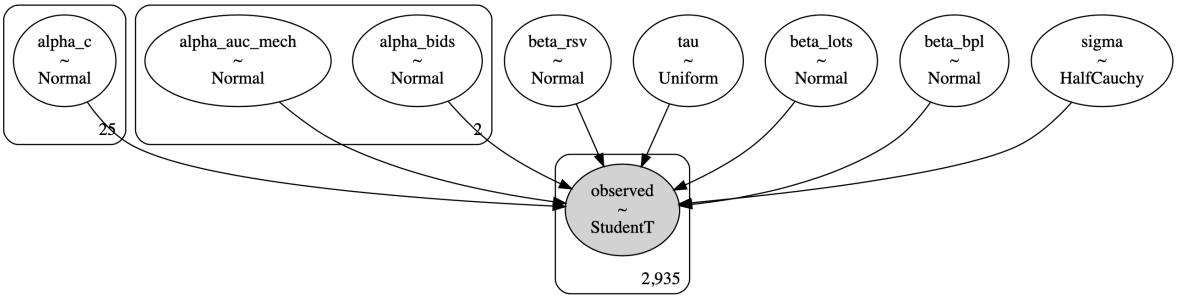

Fig 10. Graph model of model 3

When printing out the model summary we found that r_hat values in all parameters was 1, mean effective sample size was large and also in plotting the trace we found that there was no divergence. These all prove that our model is acceptable. When I plotted the forest plot I found that spread is small as compared to model 1 which is also good sign. Let's plot the prediction plot to see if model performed well or not.

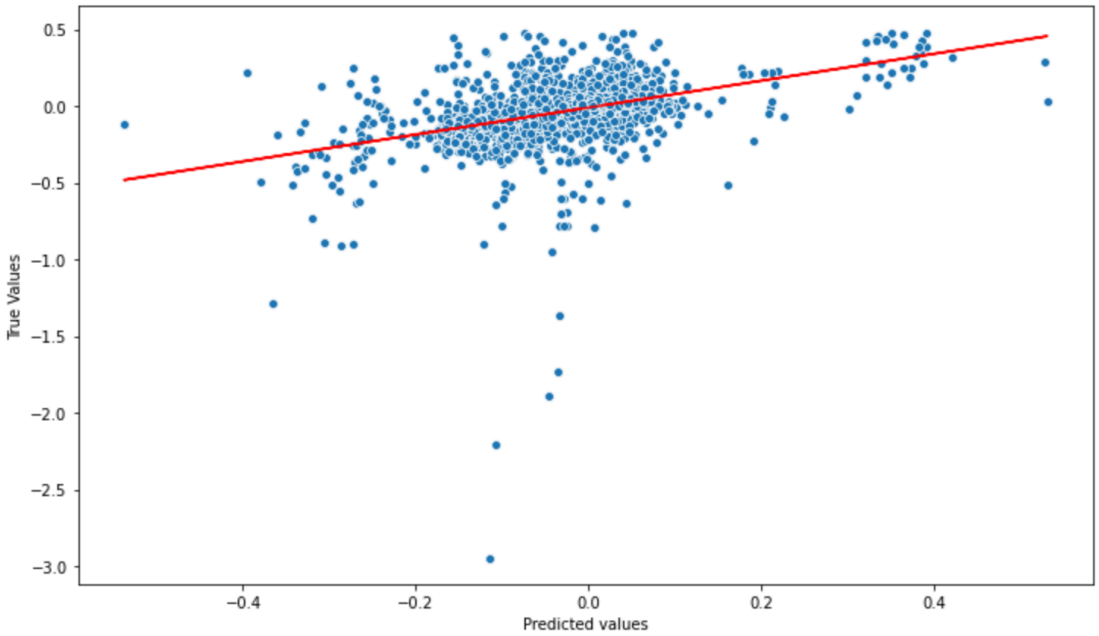

Fig 11. Model Prediction vs True values

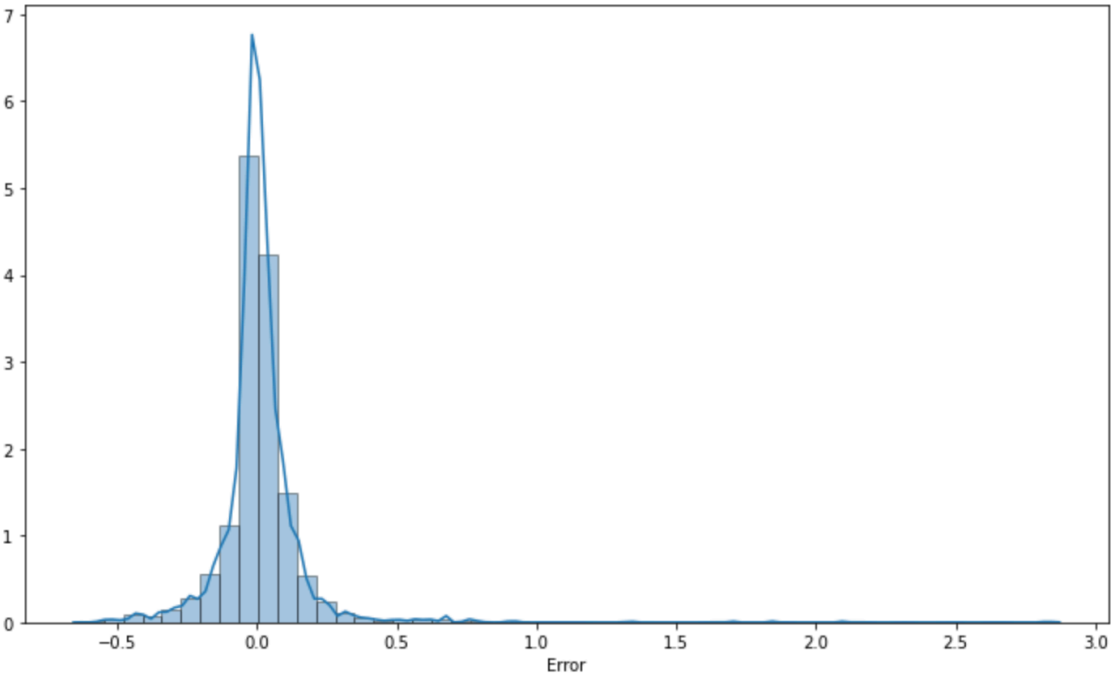

Fig 12. Residual error distribution

As we can see that model this ok job here with some bad predictions. Also residual plot looks like a nice student t-distribution.

MODEL 4

I used hierarchical model[7] with same parameters as model 2 or the hyper priors. For this model the draw size was less because my machine was not handling large draw and burn in sample size.

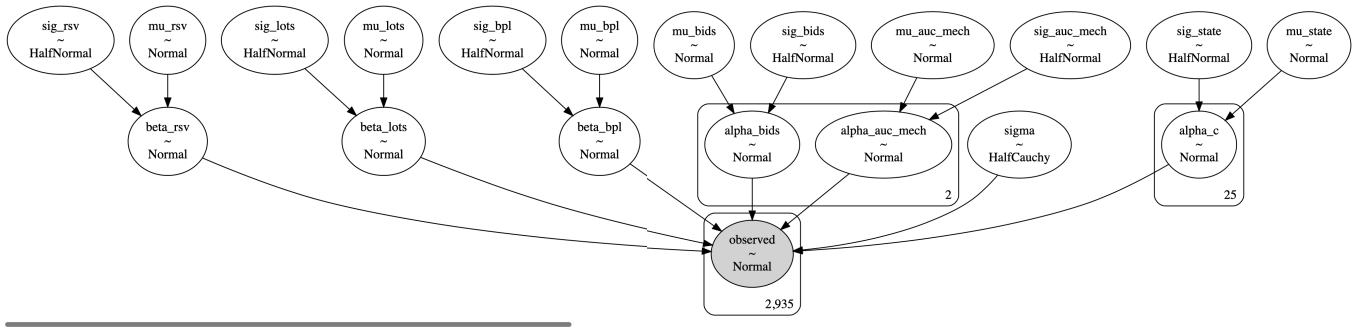

Fig 13. Graph model of Hierarchical model

When printing out the model summary we found that r_hat values in all parameters was 1, mean effective sample size was large and also in plotting the trace we found that there was no divergence. These all prove that our model is acceptable. When I plotted the forest plot I found that spread is kinda large. Let's plot the prediction plot to see if model performed well or not.

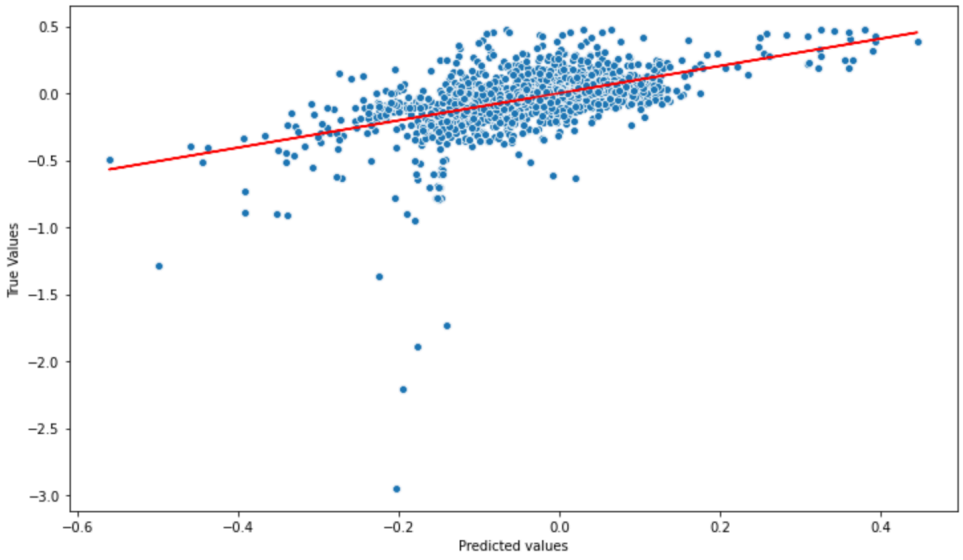

Fig 14. Model prediction vs True Values

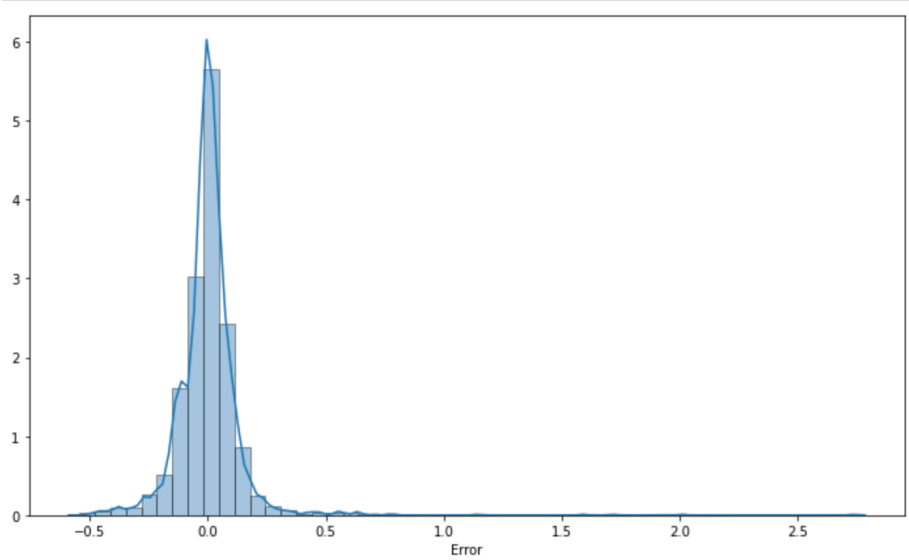

Fig 15. Residual error distribution

As we can see that from the above plot that model did ok and in there are bad predictions made by this model too. Let's compare the model to conclude which we can use for further analysis.

MODEL COMPARISON

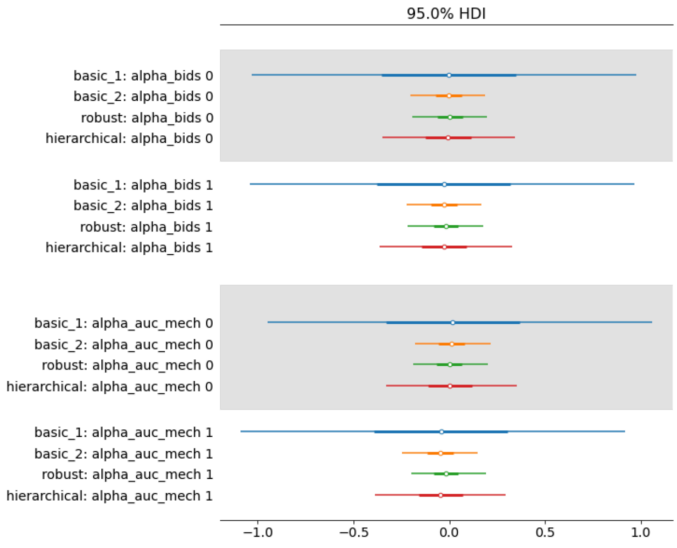

Fig 16. Forest Plot of categorical features from trace of all models

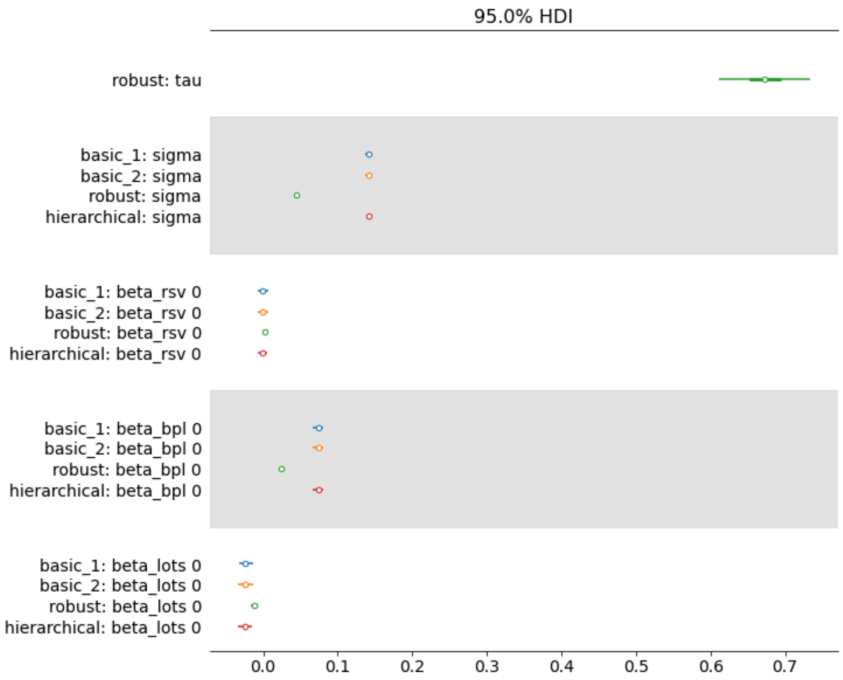

Fig 17. Forest plot of continuous features from trace of all models

As we can see from the above 2 plots that except model 1 others are mostly in line with each other. We can see that the robust model which is the model in which we used student t-distribution performed the best. Let's compare the models based on RMSE values.

MODELS	RMSE
Model 1	0.14147
Model 2	0.14062
Model 3 (with student t-distribution)	0.15079
Model 4 (Hierarchical Model)	0.14049

Table 1. All model comparison based on RMSE

As we can see that in the above table that we can't select model based on the RMSE values alone as RMSE values of our model are very close to each other. After RMSE, I used WAIC to compare the models.

	rank	loo	p_loo	d_loo	weight	se	dse	warning	loo_scale
robust	0	2810.37	52.5322	0	NaN	250.851	0	False	log
basic_1	1	1543.4	69.3566	1266.97	0	73.2747	240.905	True	log
basic_2	2	1543.04	69.0581	1267.32	0	249.613	241.142	True	log
hierarchical	3	1542.32	73.0534	1268.05	0	249.837	242.257	True	log

Table 2. WAIC comparison

As we can see that indeed model 3, the model in which we used student t-distribution performed the best, so I'm going to use model 3 for my feature analysis.

RESULTS

ANALYSIS 1 : Posterior Predictive Check

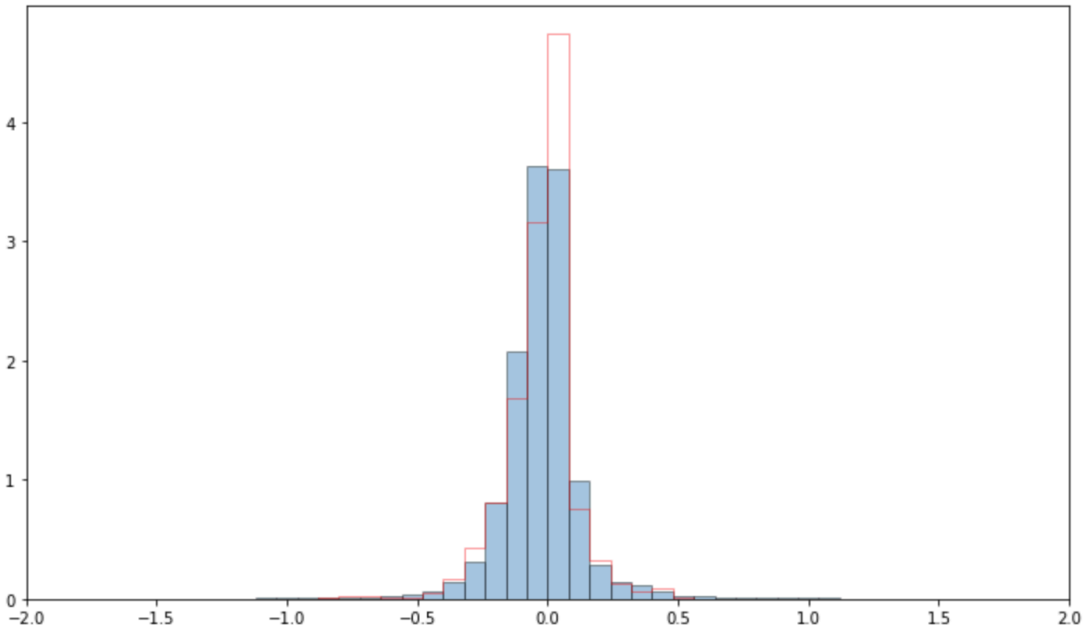

Fig 18. RoR distribution from model 3 prediction vs RoR distribution from test set

For first analysis, I plotted the distribution the RoR prediction from the model 3 and compared it with test set data (red overlay). We can say that distribution of data is more centred around the data's mean than that of model's prediction data mean.

ANALYSIS 2: Forward Auction vs Sealed-Bid

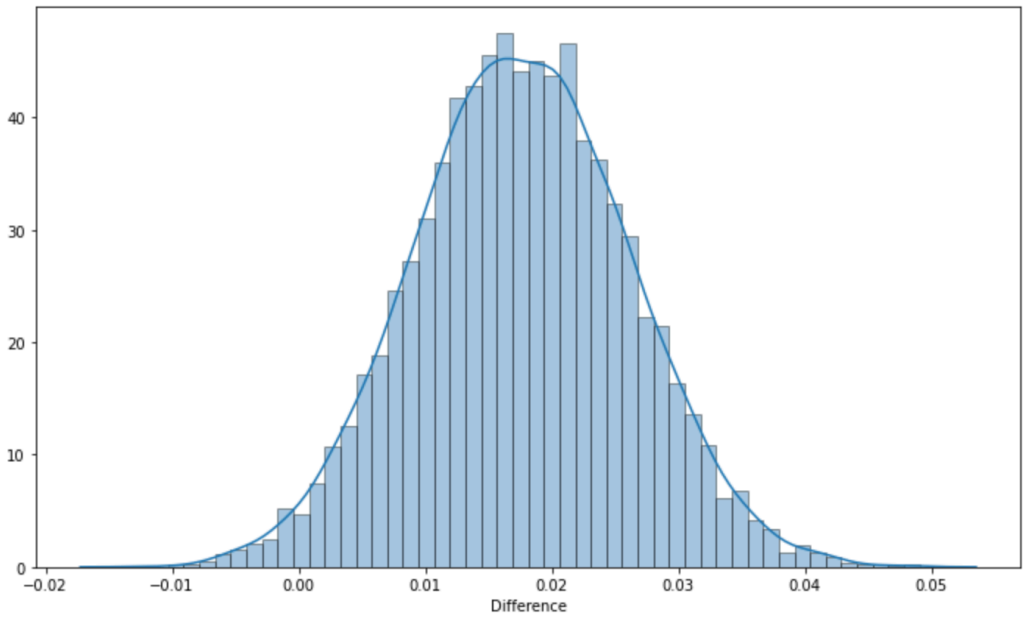

Fig 19. Distribution of difference in RoR of Forward and Sealed Bid Auction

For this analysis I used the posterior predictive values from model 3 and then I plotted the difference between the Forward auction and sealed bid. I got mean difference of 0.017613 between the two auction mechanism. We can say that Forward auction outperform the Sealed bid as they have higher RoR than sealed bid auction.

ANALYSIS 3: Impact of Bidders Per Lot and Number of Lots on ROR

For this analysis I plotted the posterior predictive values from model 3 and then plotted it against BPL feature from the dataset to see if bidders per lot have negative or positive impact on RoR.

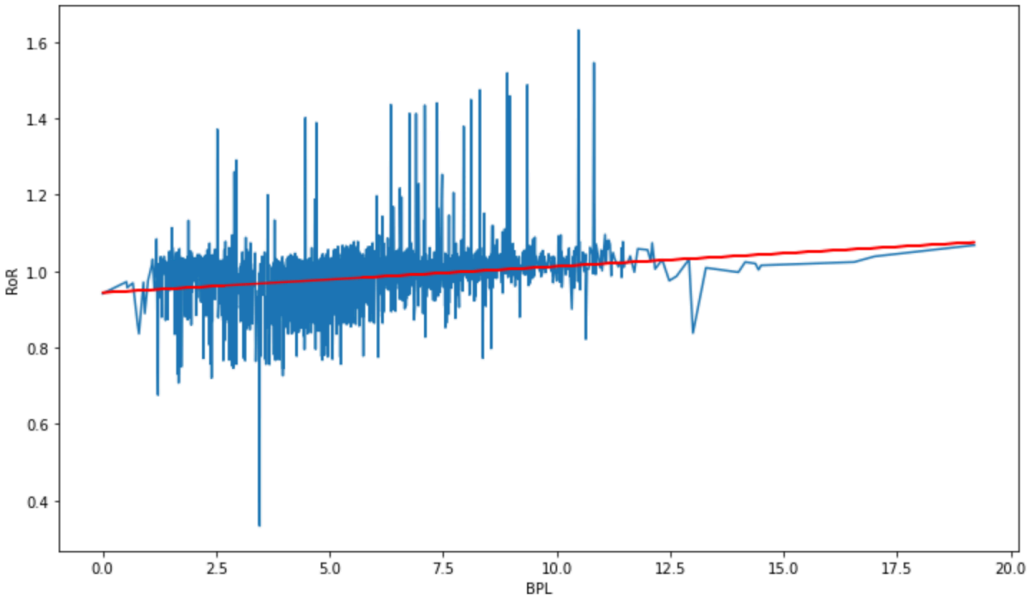

Fig 20. Bidders per lot vs RoR

As we can see in the above plot that bidders per lot have a positive impact on RoR. This was kinda expected too because the bidders are in a lot, the more RoR will be.

ANALYSIS 4: Impact of Lots offered on RoR

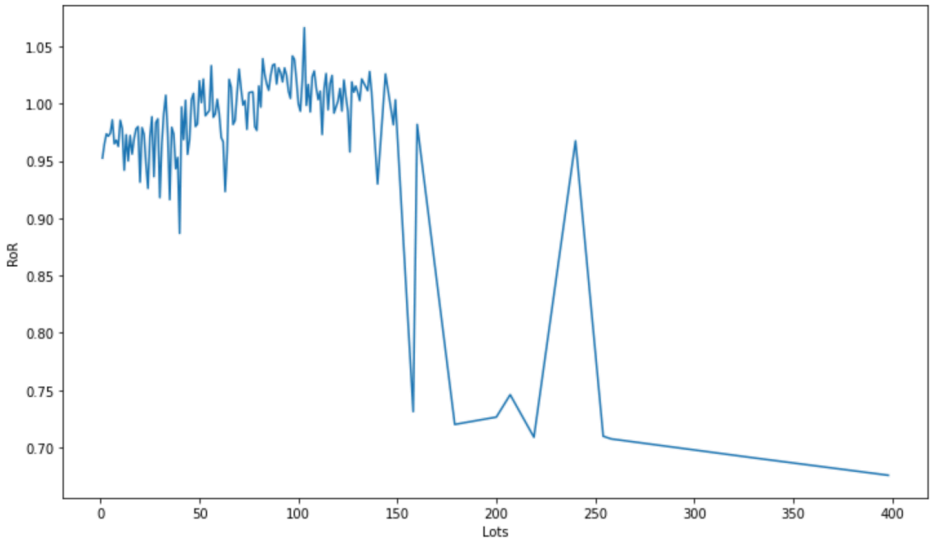

Fig 21. Lots offered vs RoR

As we can see that in the above plot that lots offered have a positive impact on ROR till 100 items in a lot but after that lots offered have negative impact on RoR. One interesting thing to notice is that there's a spike in RoR around 230 items in lot. It maybe because there's still some outliers in the data.

CONCLUSION

- On an average Forward Auction are expected to outperform Sealed Bid auctions.
- As bidders were increased, they tend to have a positive effect on RoR which is also expected that if there more bidders then RoR will be high in those cases.
- If there are more lots offered in an auction then it will have negative effect on RoR. There seem to be little positive effect on RoR when lots were small but as they increased then lots had a negative effect on RoR.

So from above results we can conclude that in order to maximise RoR from auction, auctioneer should focus more more on holding forward auctions as compared to sealed bid auctions with less lots items in a lot. Also we see that if there are more bidders per lot than it tend to have a positive impact on RoR.

BIBLIOGRAPHY

[1] Kenton, Will. "Sealed-Bid Auction." *Investopedia*, Investopedia, 29 Aug. 2020, www.investopedia.com/terms/s/sealed-bid-auction.asp

[2] Brandly, Mike, and Mike Brandly. "English Auction versus Sealed Bid and Dutch Auctions." *Mike Brandly, Auctioneer Blog*, 30 July 2012, mikebrandlyauctioneer.wordpress.com/2012/07/29/english-auction-versus-sealed-bid-and-dutch-auctions/

[3] Lacetera, Nicola, et al. Bid Takers or Market Makers? The Effect of Auctioneers on Auction Outcome. 2016, web.stanford.edu/~bjlarsen/LLPS_2016.pdf

[4] Jason Stoughton October 23, et al. "'The Greatest Auction Ever' – Q&A with Paul Milgrom, 2020 Nobel Laureate." *Beta Site for NSF - National Science Foundation*, 23 Oct. 2020, beta.nsf.gov/science-matters/greatest-auction-ever-qa-paul-milgrom-2020-nobel-laureate

[5] Milgrom, Paul, and Ilya Segal. "Clock Auctions and Radio Spectrum Reallocation." *Journal of Political Economy*, vol. 128, no. 1, 2020, pp. 1–31., doi:10.1086/704074

[6] Salvatier1, John, et al. "Probabilistic Programming in Python Using PyMC3." *PeerJ Computer Science*, PeerJ Inc., 6 Apr. 2016, peerj.com/articles/cs-55/.

[7] Elbers, Danne. "GLM: Hierarchical Linear Regression." GLM: Hierarchical Linear Regression - PyMC3 3.9.3 Documentation, 2016, docs.pymc.io/notebooks/GLM-hierarchical.html

[8] Wiecki, Thomas. "While My MCMC Gently Samples." *While My MCMC Gently Samples Atom*, twiecki.io/blog/2013/08/12/bayesian-glms-1/

[9] Shah, Amar, et al. "Student-t Processes as Alternatives to Gaussian Processes." *ArXiv.org*, 19 Feb. 2014, arxiv.org/abs/1402.4306

[10] Levin, Jonathan. *Auction Theory*. Oct. 2004, web.stanford.edu/~jdlevin/Econ%20286/Auctions.pdf