INSURANCE CLAIM

Prediction & Classification

ARUN RAJESH B.

INDIAN INSTITUTE OF TECHNOLOGY MADRAS • सिद्धिर्भवति कर्मजा •

AGENDA

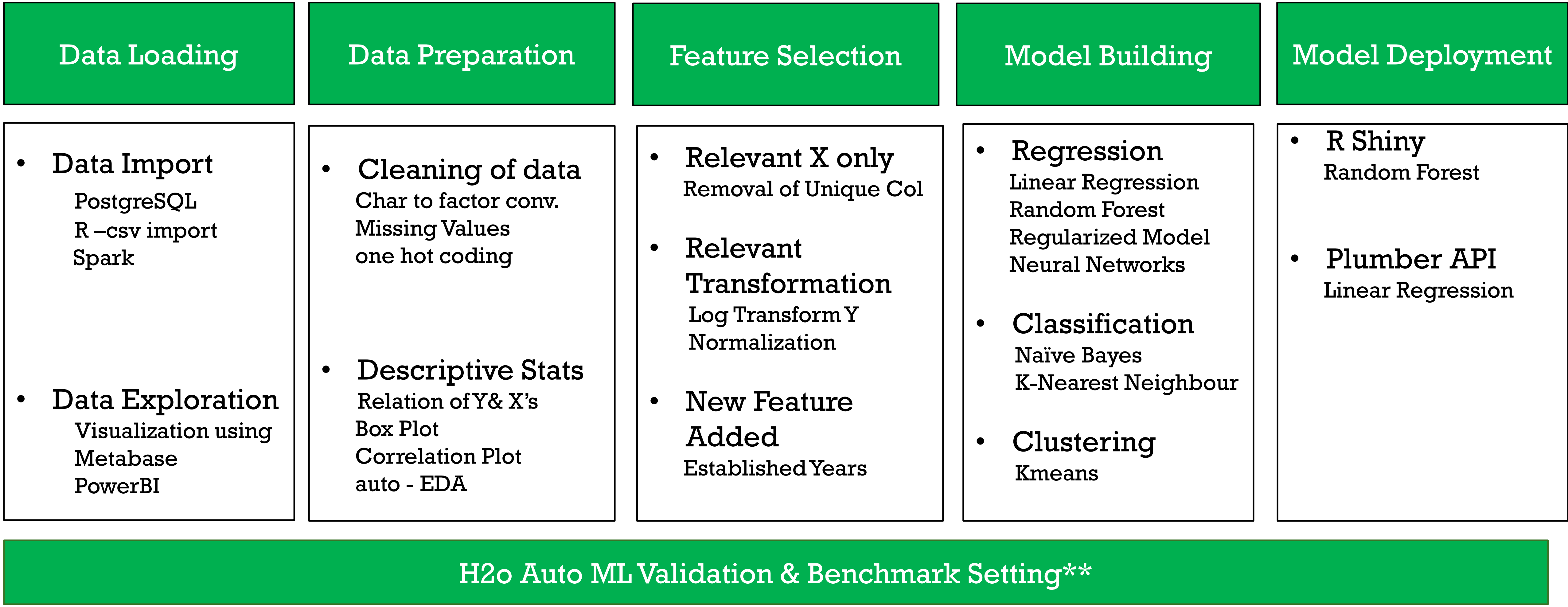
1. Project Background & Objectives
2. Project Framework
3. Business Insights– Metabase and PowerBi
4. ML - Prediction Results
5. ML - Classification Results
6. ML - Clustering Results
7. Model deployment – Shiny app and Plumber API

Project Background & Objectives

Insurance Prediction dataset has 1.8L observations – 6 Numeric columns and 12 factor columns.

1. Explore the data & deliver key business Insights
2. Fit a regression model with highest adjusted R-square and least RMSE
3. Perform classification on the same dataset & achieve high accuracy scores
4. Perform clustering on the dataset and arrive at optimal clusters
5. Deploy the model for production ready

Project Approach/Framework



Data Loading

- Data Import
  PostgreSQL
  R –csv import
  Spark
- Data Exploration
  Visualization using
  Metabase
  PowerBI

Data Preparation

- Cleaning of data
  Char to factor conv.
  Missing Values
  one hot coding
- Descriptive Stats
  Relation of Y& X's
  Box Plot
  Correlation Plot
  auto - EDA

Feature Selection

- Relevant X only
  Removal of Unique Col
- Relevant Transformation
  Log Transform Y
  Normalization
- New Feature Added
  Established Years

Model Building

- Regression
  Linear Regression
  Random Forest
  Regularized Model
  Neural Networks
- Classification
  Naïve Bayes
  K-Nearest Neighbour
- Clustering
  Kmeans

Model Deployment

- R Shiny
  Random Forest
- Plumber API
  Linear Regression

H2o Auto ML Validation & Benchmark Setting**

**

H2o Flow: Distributed Random Forest Model : R2: 0.998, RMSE: 3231.59
H2o Flow: Gradient Boosting Machine Model : R2:0.945, RMSE: 18,801

Key Business Insights

- Average amount withdrawn under Polo4 for is highest mostly coming from semi urban region and urban regions in popular multi-speciality hospitals

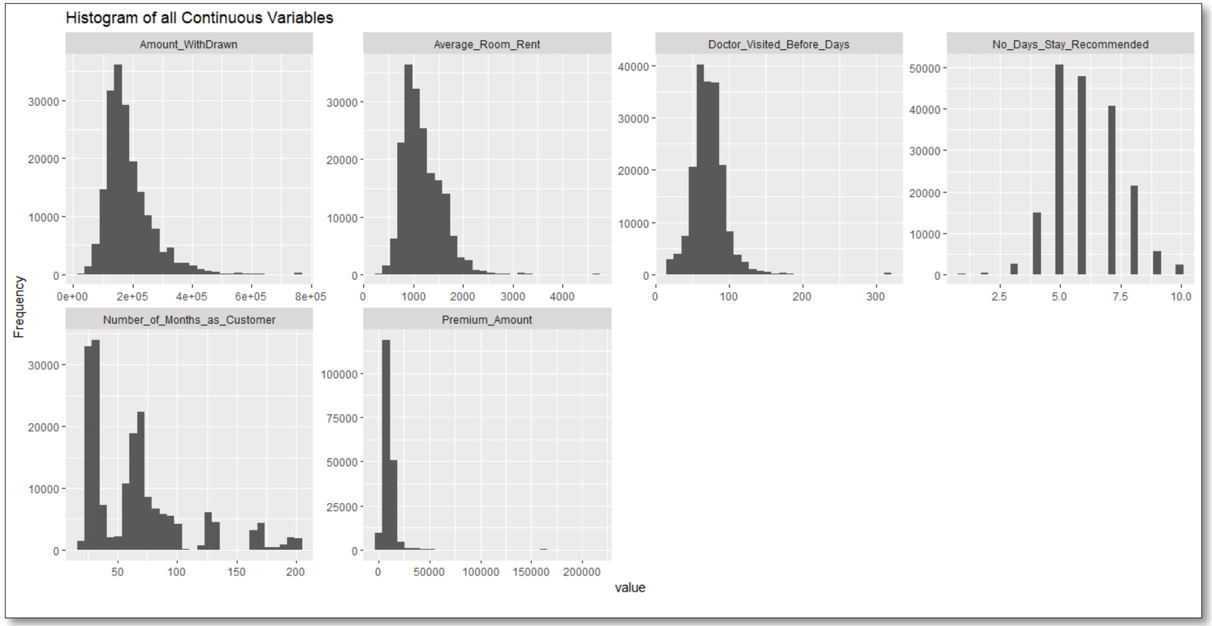
- Most withdrawals are made in Urban, followed by town (<50% of Urban), semi urban and bigger town are very less withdrawals **(scope for expansion in semi urban and bigger towns)**

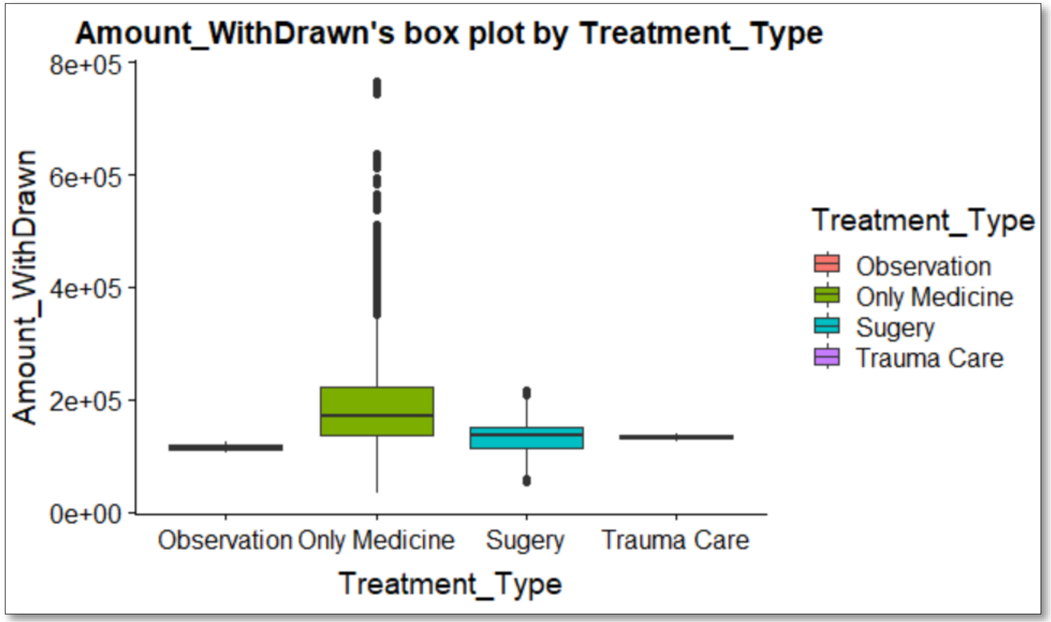
- Across all policy types, most withdrawals were made for medicine treatment type. For POLO5, comparatively higher withdrawal made for surgery treatment type

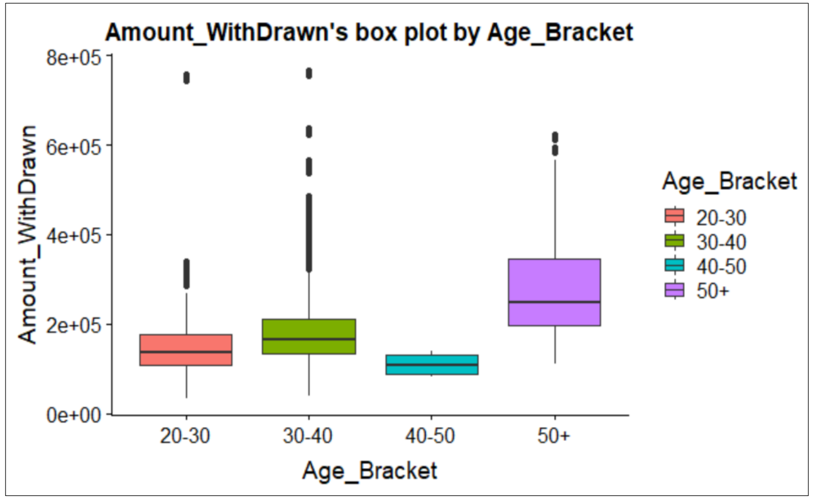
- Average amount withdrawn is highest for patients admitted under neurology followed by cardiology and endocrinology **(Scope for more hospitals to come / Health awareness to be made on neurology related disorders)**

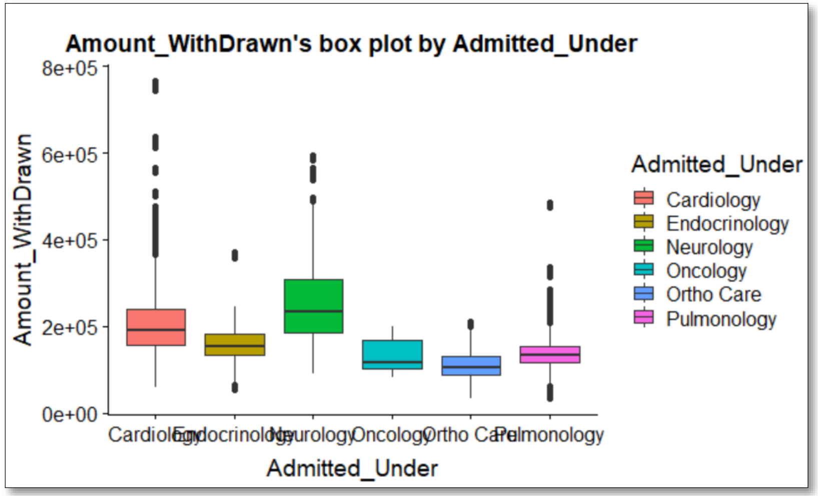
- Most withdrawals are made by people in age bracket **30 to 40 (surprising as not old)** across all hospital locations. 50 + withdraw most in bigger towns. 40 to 50 is the only age bracket which is making withdrawals for surgery and trauma care. People in other brackets are only doing it for Medicine. Average amount paid is highest for 50+ group for "medicine" treatment type.

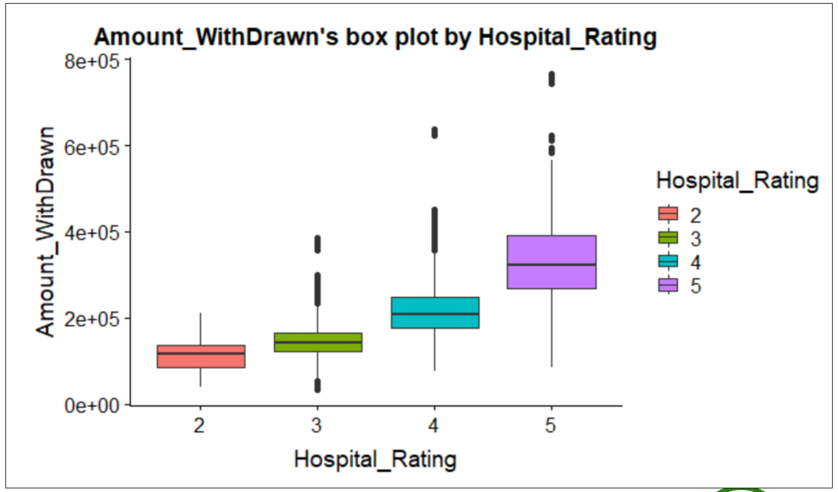
Few Exploratory analysis in R (auto – EDA)





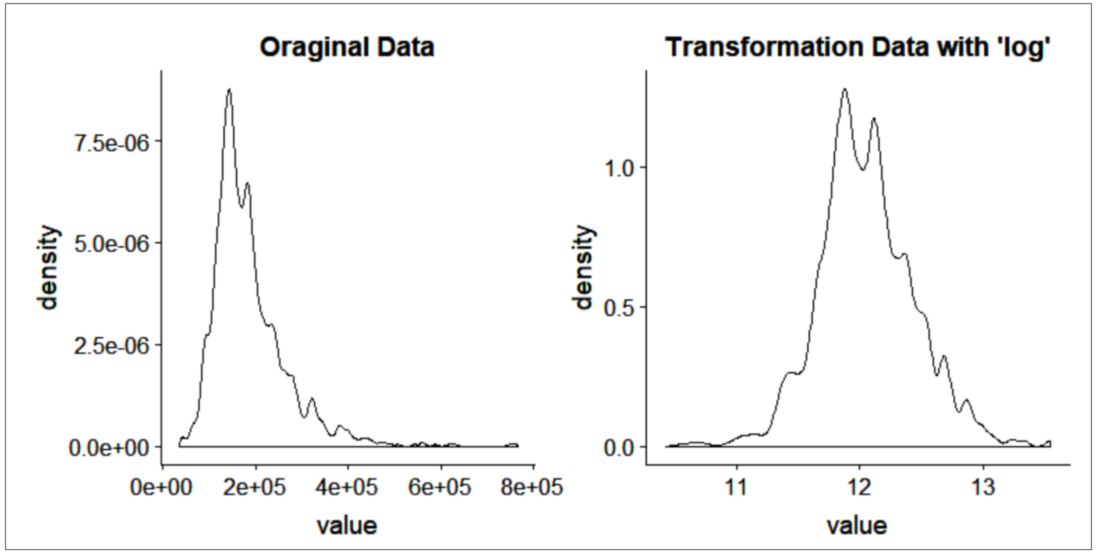






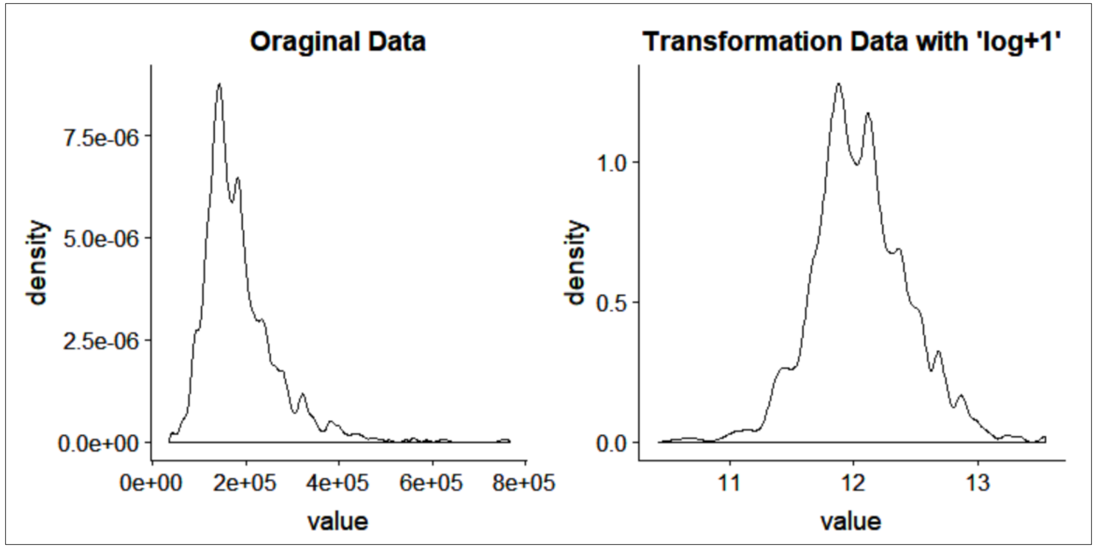
Few Transformations(auto – EDA)

Target Variable : Amount With Drawn



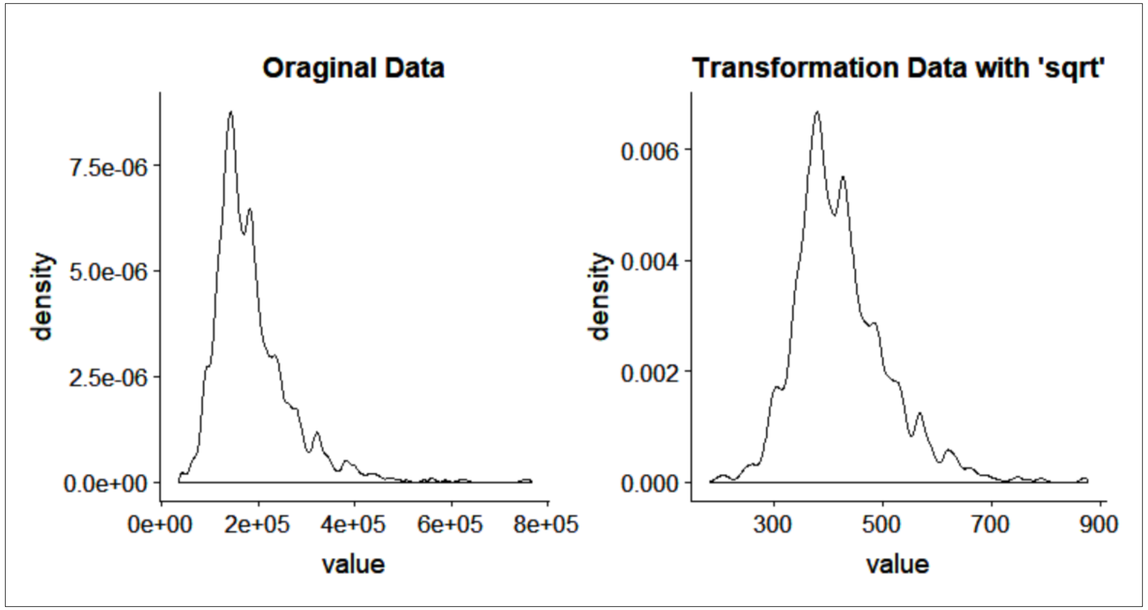
Oraginal Data

Transformation Data with 'log'



Oraginal Data

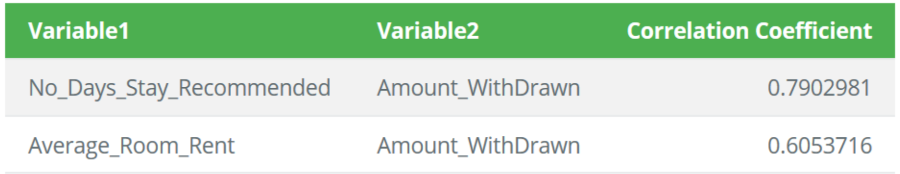
Transformation Data with 'log+1'



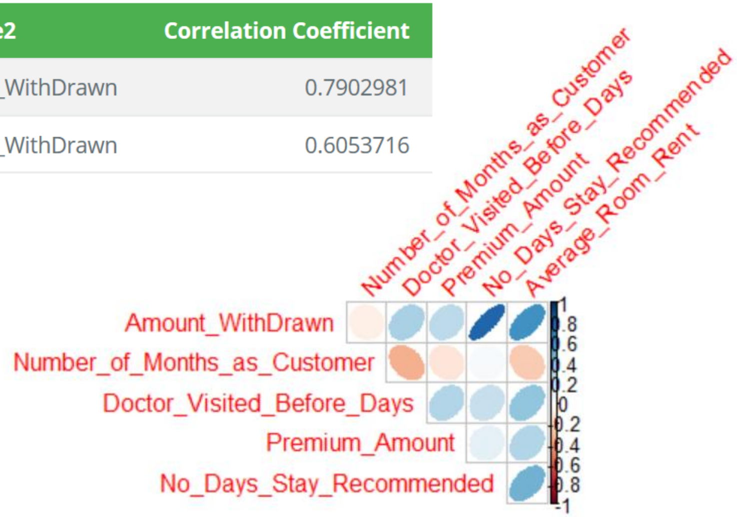
Oraginal Data

Transformation Data with 'sqrt'

Table of correlation coefficients (0.5 or more)



Variable1	Variable2	Correlation Coefficient
No_Days_Stay_Recommended	Amount_WithDrawn	0.7902981
Average_Room_Rent	Amount_WithDrawn	0.6053716



Regression Algorithm Results I

Target Variable : Amount With Drawn

S.No	Regression Algorithm	Input Pre-processing & Other	Train Model Accuracy (Percentage)	Hyper Parameter &Others	Test Model Accuracy (Percentage)
1	Linear Regression (Base Model)	- Scaling , With/Without Cross Validation	Res.SE : 0.4785 R -squared: 0.7703 (adjusted)	- CSV file output Generated for test data	RMSE : 0.4738 R-squared : 0.7740 MAE : 0.3178 MAPE : 4.1097
2	Linear Regression (Log Transform of Y Variable)	- Skewness was observed in Y Parameter. - Log Transform, Scaling , - With/Without Cross Validation	Res.SE : 0.1645 R -squared: 0.8187 (adjusted)	- CSV file output Generated for test data	RMSE : 0.1629 R-squared : 0.824 MAE : 0.1218 MAPE : 0.0101
	Linear Regression (Log Transform of Y and Continuous X Variables)	- Skewness was observed in Y Parameter. - Log Transform, Scaling , - With/Without Cross Validation	Res.SE : 0.1603 R -squared: 0.8278 (adjusted)		RMSE : 0.1599 R-squared : 0.8305 MAE : 0.1183 MAPE : 0.0098

Learnings

Log Transformation of Skewed data gives better results

Regression Algorithm Results II

Target Variable : Amount With Drawn

S.No	Regression Shrinkage Models	Input Pre-processing & Other	Train Model Performance (Percentage)	Hyper Parameter &Others	Test Model Accuracy (Percentage)
4	Lasso Regression	- Scaling ,Center - With Cross Validation	R -squared: 0.8264	- Model best tune Fraction : 0.9	RMSE : 0.1606 R-squared : 0.8294 MAE : 0.1185 MAPE : 0.0098
5	Ridge Regression	- Scaling , Center - With Cross Validation	R -squared: 0.8276	- Model best tune lambda : 0	RMSE : 0.1599 R-squared : 0.8305 MAE : 0.1183 MAPE : 0.0098
6	Elastic Net Regression	- Scaling , Center - With Cross Validation	R -squared: 0.8276	- Model best tune lambda : 0 Fraction : 1	RMSE : 0.1599 R-squared : 0.8305 MAE : 0.1183 MAPE : 0.0098

Learnings

Shrinkage Models didn't improve model performance. They yield the same result as normal Linear models for this dataset

Regression Algorithm Results III

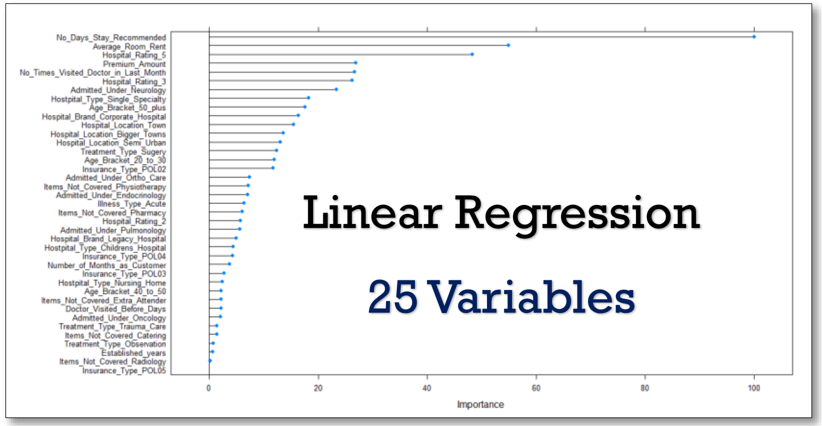
Target Variable : Amount With Drawn

S.No	Regression - Tree Based Models	Input Pre-processing & Other	Train Model Performance (Percentage)	Hyper Parameter &Others	Test Model Accuracy (Percentage)
7	Random Forest (Spark – ML Lib through Sparklyr Package)	- Log Transform of Y Parameter	RMSE: 0.05376 R2:0.9805	Num. trees =80 Max_depth =12	RMSE : 0.0537 R-squared : 0.9811 MAE : 0.0376 MAPE : 0.0032
8	Random Forest (Ranger Package)	- Scaling , - With Cross Validation	OOB prediction error (MSE): 0.01809 R squared (OOB): 0.98184	- Num. trees =500 - Mtry =6 - Node. Size =5	RMSE : 0.1377 R-squared : 0.9851 MAE : 0.0935 MAPE : 1.3804
	Random Forest (Ranger Package)	- No Scaling of X Parameters - Log Transform of Y Parameter	OOB prediction error (MSE): 0.00048 R squared (OOB): 0.99677	- Mtry= 11:17 - Min. Node. Size = 3:9 - Splitrule="variance“ - CSV file output Generated Optimum Value Mtry =17 Min. Node. Size =9	RMSE : 0.2212 R-squared : 0.9967 MAE : 0.0167 MAPE : 0.0014

Learnings

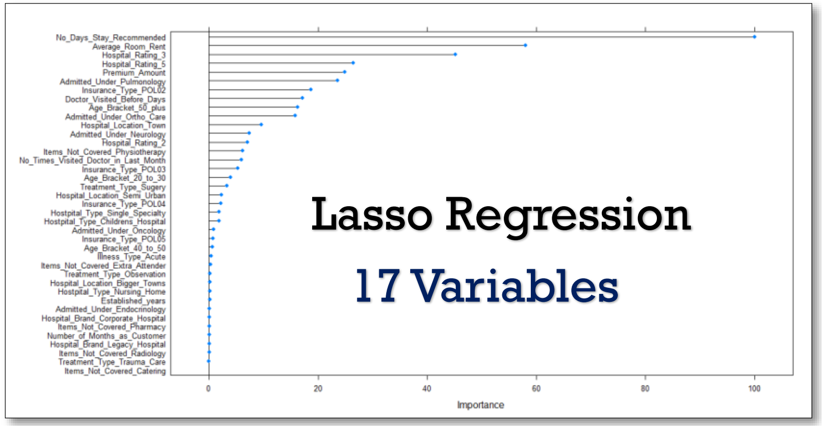
Tree models perform better than regression models when hyperparameter tuning is involved

Variable Importance & Model Interpretation (LIME)



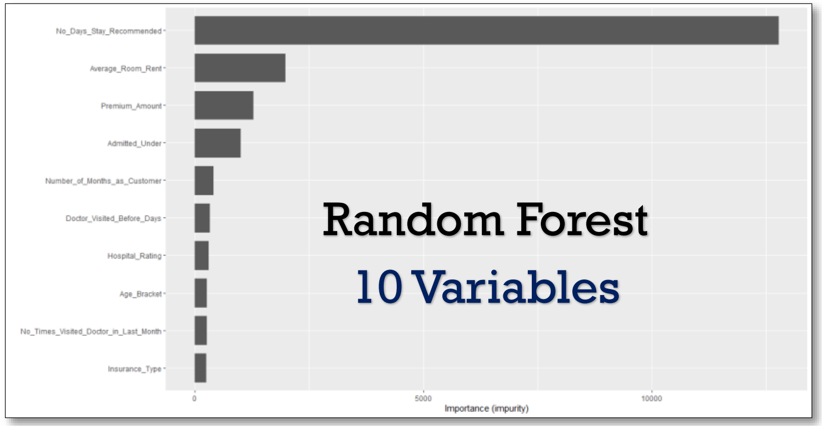
Linear Regression

25 Variables



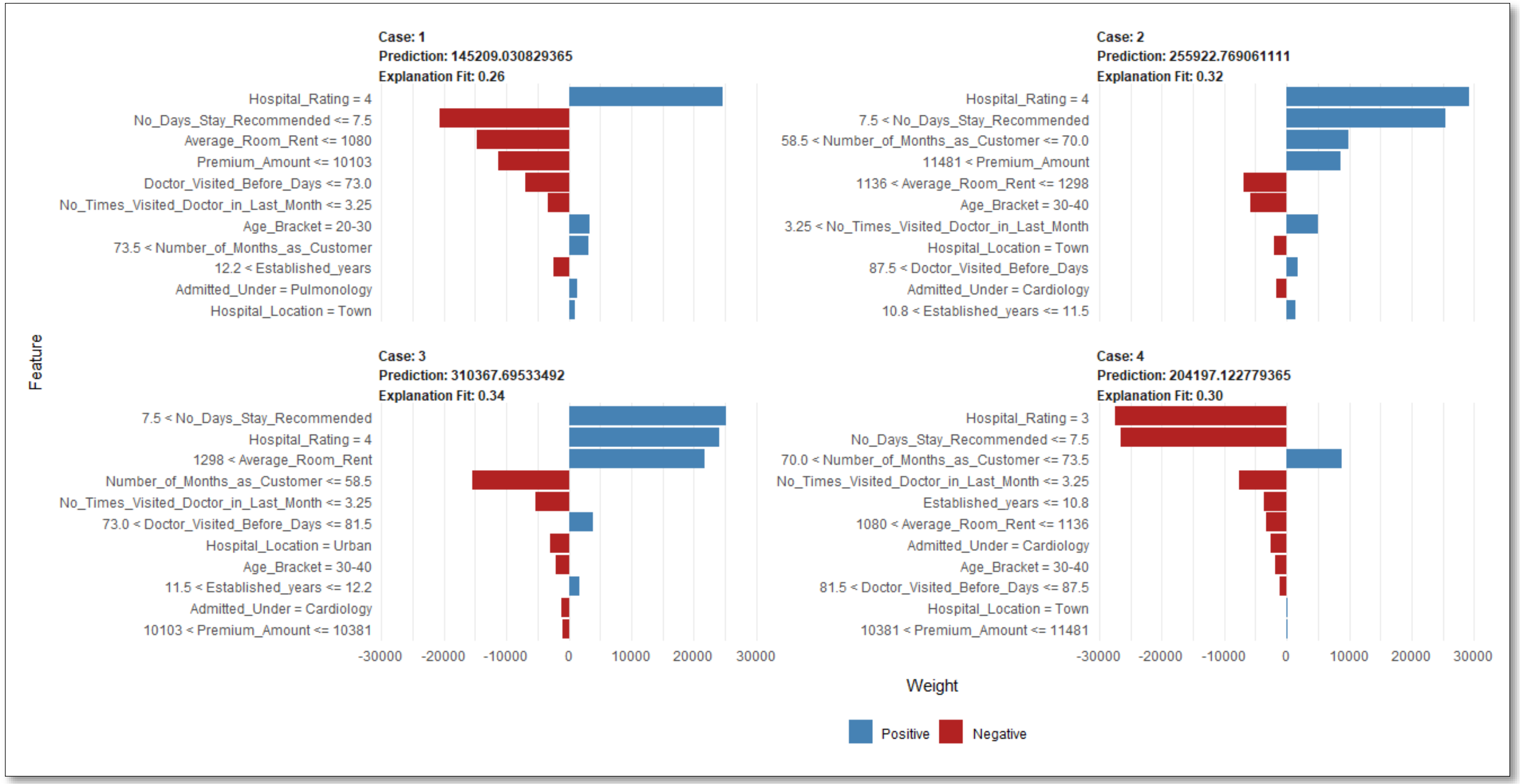
Lasso Regression

17 Variables



Random Forest

10 Variables

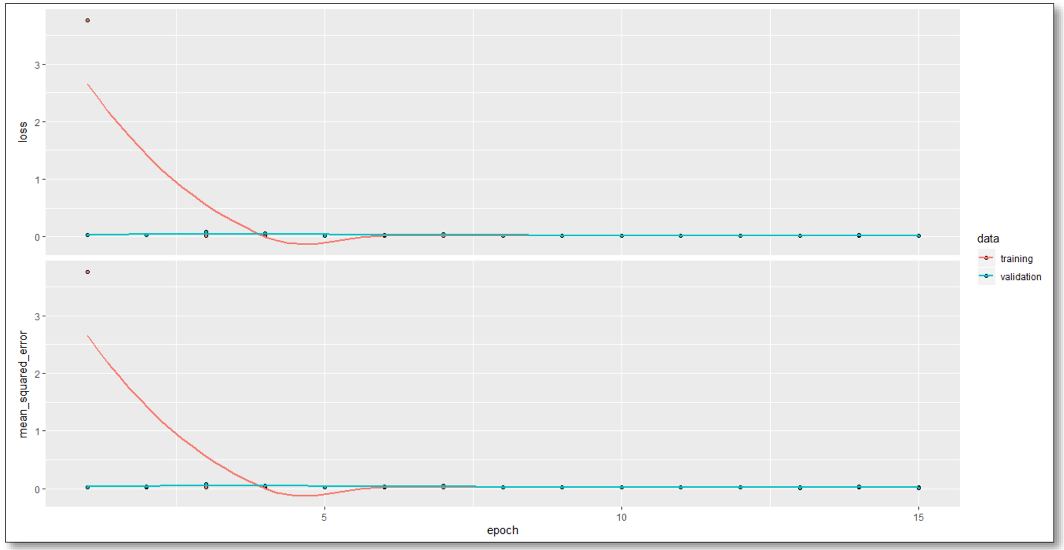


- local interpretations of ML models using 4 Observations of training dataset (LIME Package)
- 11 influential variables are plotted. Explains the support of variables for prediction
- No. of Days stay recommended, Hospital Rating, Average Room Rent are the most influential variables

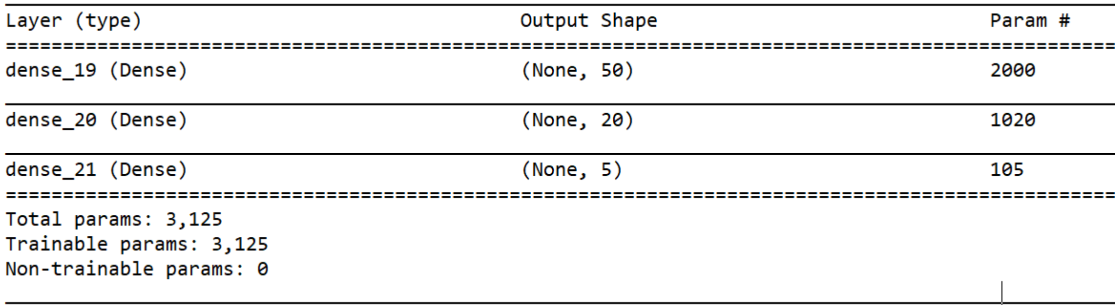
Regression Algorithm Results IV

Target Variable : Amount With Drawn

S.No	Regression - Neural Networks	Input Pre-processing & Other	Train Model Performance (Percentage)	Hyper Parameter &Others	Test Model Accuracy (Percentage)
9	Neural Networks – Keras Package	- Log Transform of Y Parameter - Scaling, Center values of X Parameter	Train loss: 0.0175 – mean_squared_error: 0.0175 Validation loss: 0.0135 mean_squared_error: 0.0135	Compile function: optimizer = "rmsprop", loss = "mse", metrics = c("mse") Training epochs = 15, batch_size = 32, validation_split = 0.2	RMSE : 0.1215 R-squared : 0.9223 MAE : 0.0954 MAPE : 0.0079



# network architecture
layer_dense(units = 50, activation = "relu", input_shape = ncol(train_x)) %>%
layer_dense(units=20,activation="relu")%>%
layer_dense(units = 5, activation = "relu")



```
Layer (type)                          Output Shape                    Param #
=================================================================================
dense_19 (Dense)                      (None, 50)                      2000
_________________________________________________________________________________
dense_20 (Dense)                      (None, 20)                      1020
_________________________________________________________________________________
dense_21 (Dense)                      (None, 5)                       105
=================================================================================
Total params: 3,125
Trainable params: 3,125
Non-trainable params: 0
_________________________________________________________________________________
```

Classification Algorithm Results

Target Variable : Amount With Drawn
Class: High, Low

S.No	Classification Algorithm	Input Pre-processing & Other	Train Model Accuracy (Percentage)	Hyper Parameter Condition & Optimum value	Test Model Accuracy (Percentage)
1	Naïve Bayes Classification algorithm	- 3 Fold Repeated Cross Validation - No scaling	90.05	Laplace=1 to 5, usekernel=TRUE,FALSE, adjust=1 to 5 Optimum Values: Laplace = 1.5, Usekernel=TRUE Adjust=2	Accuracy : 90.02 Sensitivity : 88.84 Specificity : 91.20
2	K-Nearest Neighbours algorithm.	- Without Scaling		K = 7 to 17 Optimum Value: K = 11	Accuracy : 98.23 Sensitivity : 98.30 Specificity : 98.17
		- With Scaling (Note: Scaling execution was little faster than the other)		K = 7 to 17 Optimum Value: K = 11	Accuracy : 98.48 Sensitivity : 98.46 Specificity : 98.50

Learnings

Many Zero Probabilities were identified during execution of Naïve Bayes algorithm. Used Laplace smoothing to prevent such errors

Clustering Algorithm Results

S.No	Clustering Algorithm	Input Pre-processing & Other	Hyper Parameter Condition & Optimum value	Observations

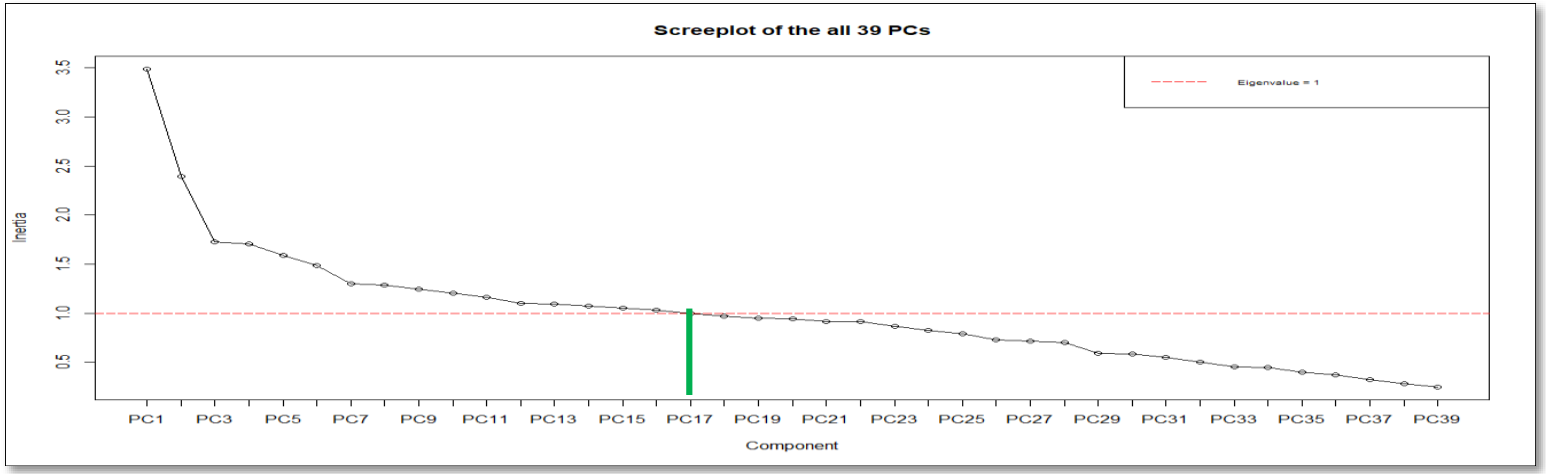
1	K means Clustering	- PCA performed on input parameters to reduce the dimensions
		- Base on the screeplot, Dimension from PC1 to PC 17 are selected and clustering is performed
			Kmeans Parameters
			K = 1 to 30
			Max.Iter = 100
			Algorithm = MacQueen
				Though, PC1 to PC17 is chosen from Screeplot (eigen value >1), the total variation explained is only 63%
				Optimum no. of clusters couldn't be established using Elbow Method.
				Attempt is made to plot 10 cluster on a 2 dimensional space

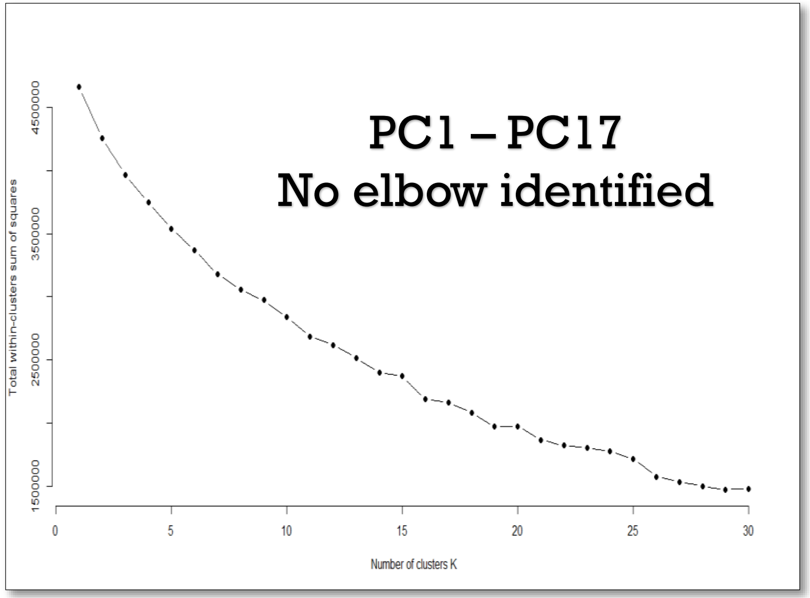
2	K Means Clustering – ClusterR Package	- PCA performed and dimensions are reduced to 2	max_clusters =15
			criterion = WCSSE
			max_iters =100
			initializer = kmeans++
				PC1 and PC2 explains only 15% variation in the original data
				Sharp decline observed at cluster 3 and 5.
				Attempt is made to plot 5 cluster on a 2 dimensional space

Learnings

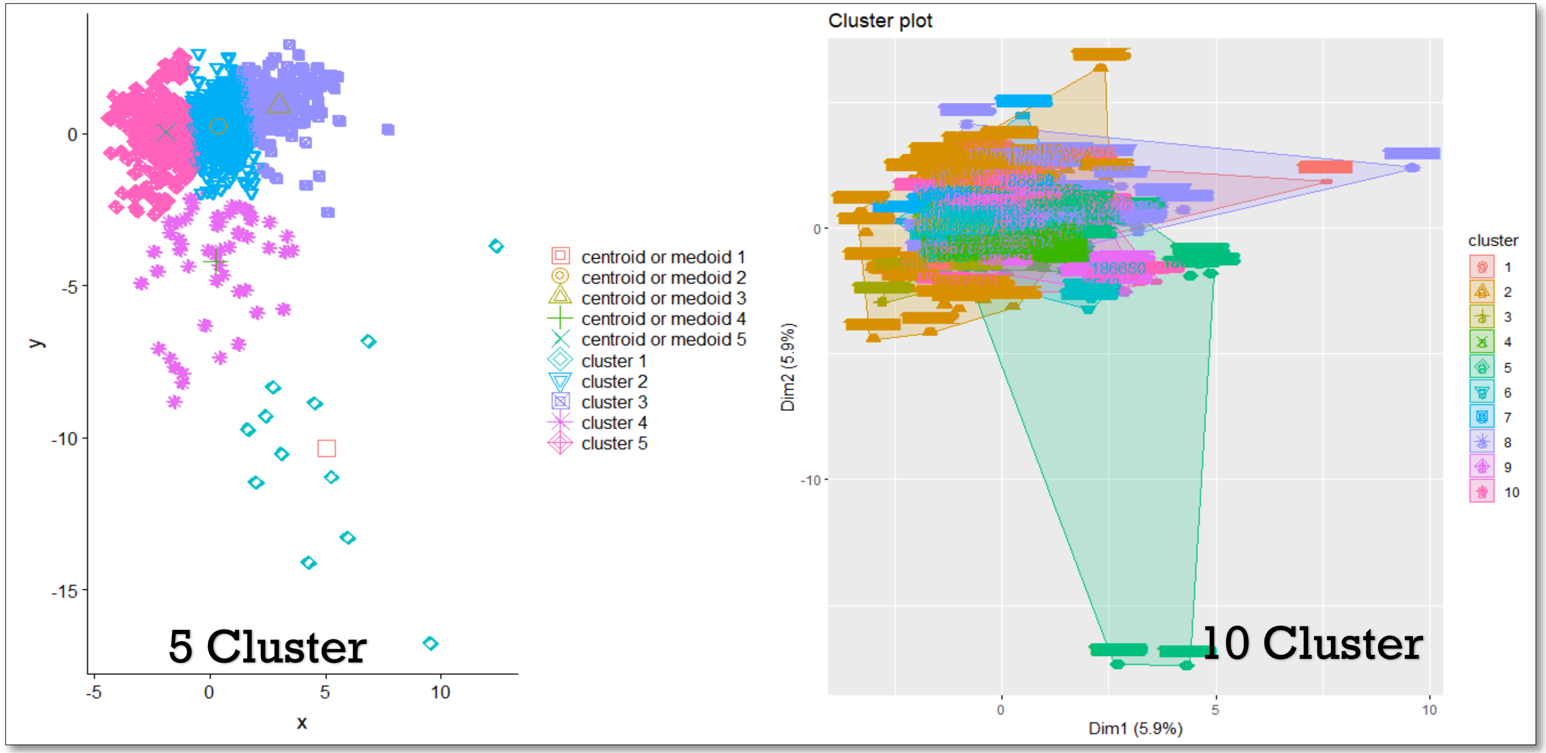
Optimum cluster based on elbow method couldn't be figured out. The total variation explained by PCA is less

Clustering Algorithm Results



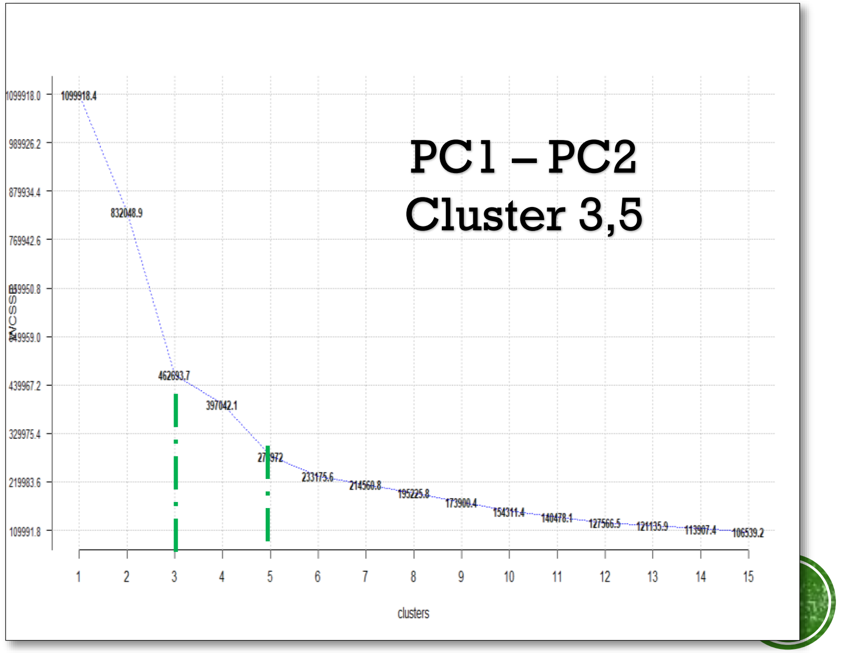


PC1 – PC17
No elbow identified



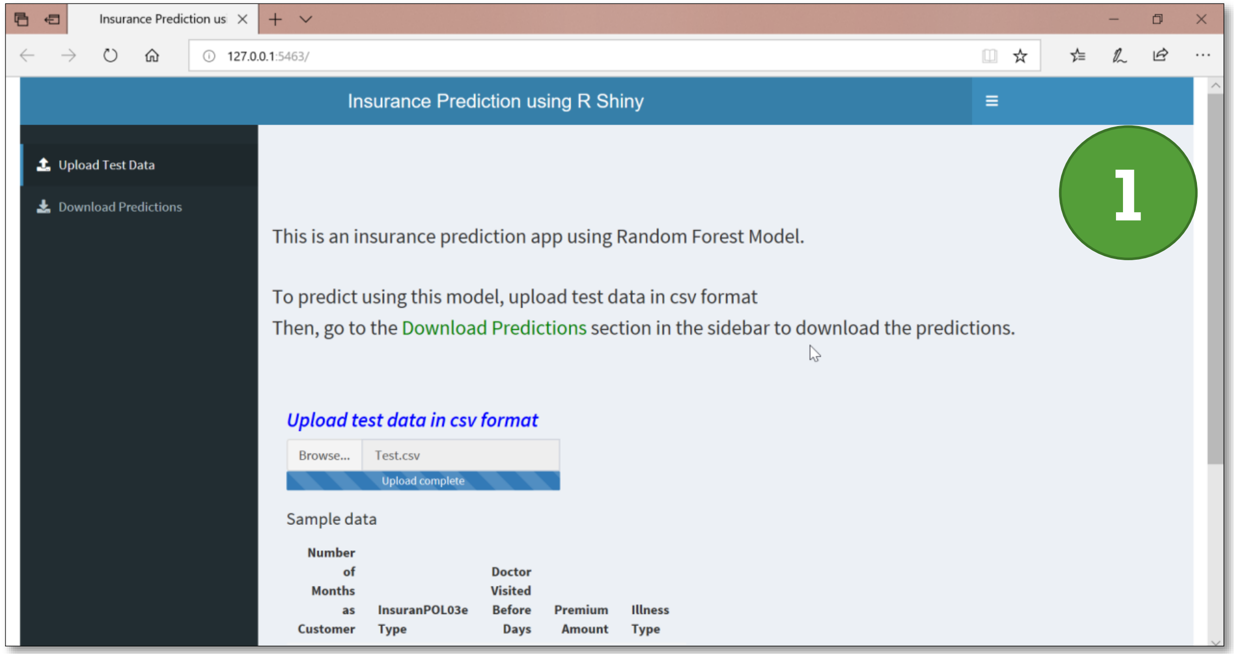
5 Cluster

10 Cluster



PC1 – PC2
Cluster 3,5

Model Deployment – R Shiny



Insurance Prediction using R Shiny

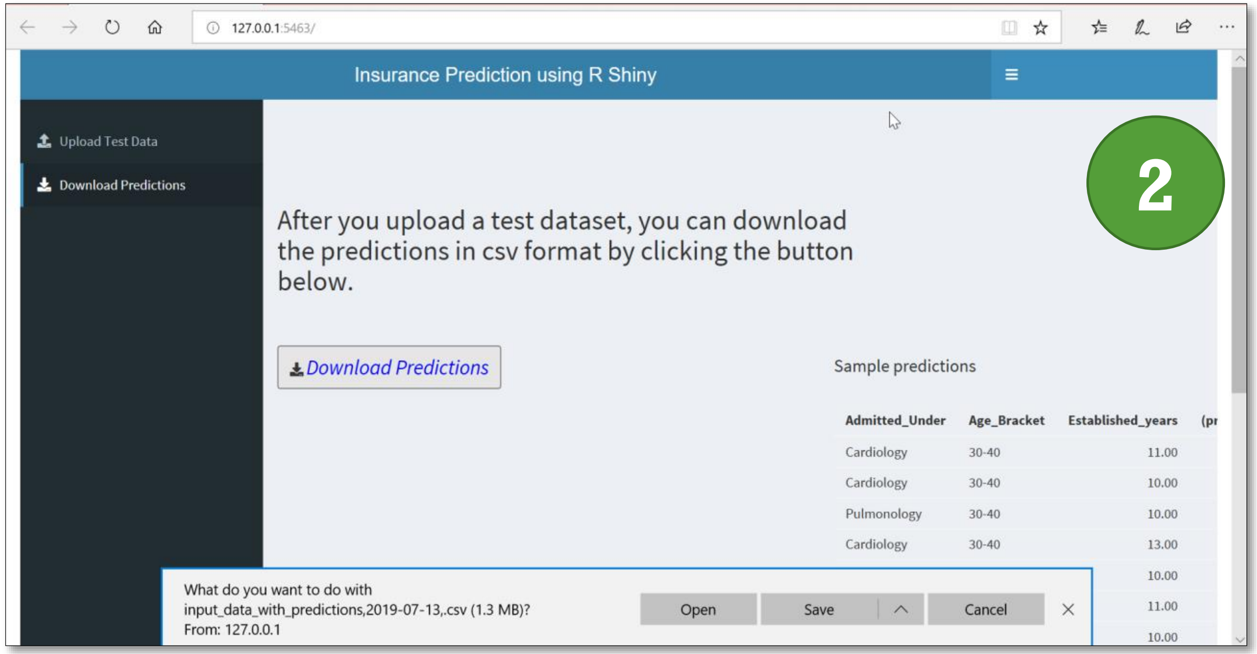
This is an insurance prediction app using Random Forest Model.

To predict using this model, upload test data in csv format
Then, go to the Download Predictions section in the sidebar to download the predictions.

Upload test data in csv format

Sample data

Number of Months as Customer	InsuranPOL03e Type	Doctor Visited Before Days	Premium Amount	Illness Type



Insurance Prediction using R Shiny

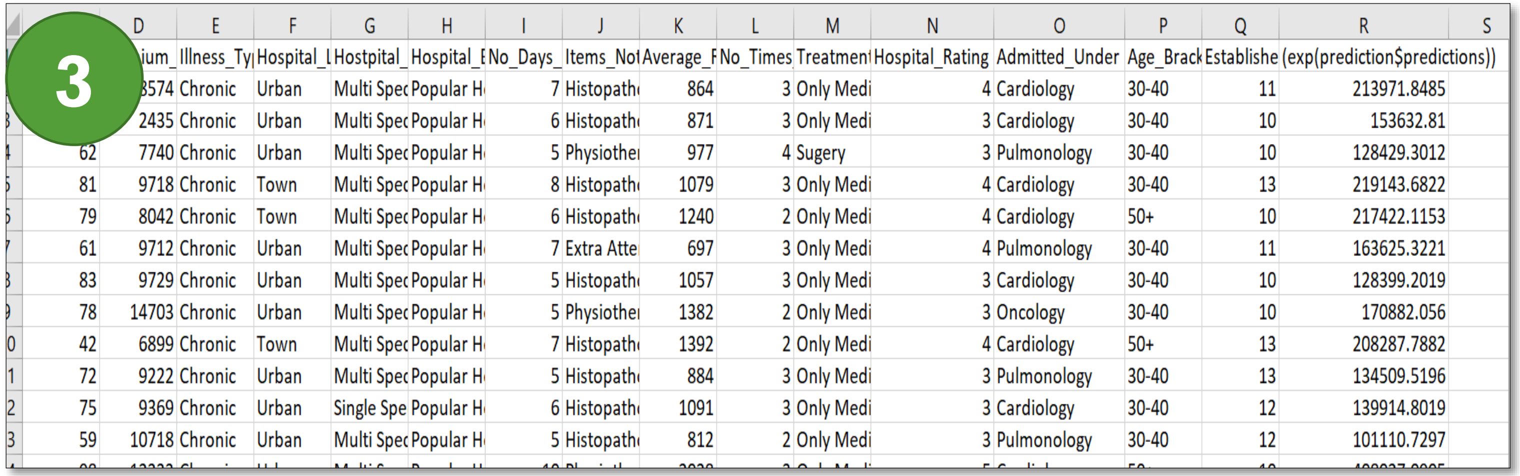
After you upload a test dataset, you can download the predictions in csv format by clicking the button below.

Download Predictions

Sample predictions

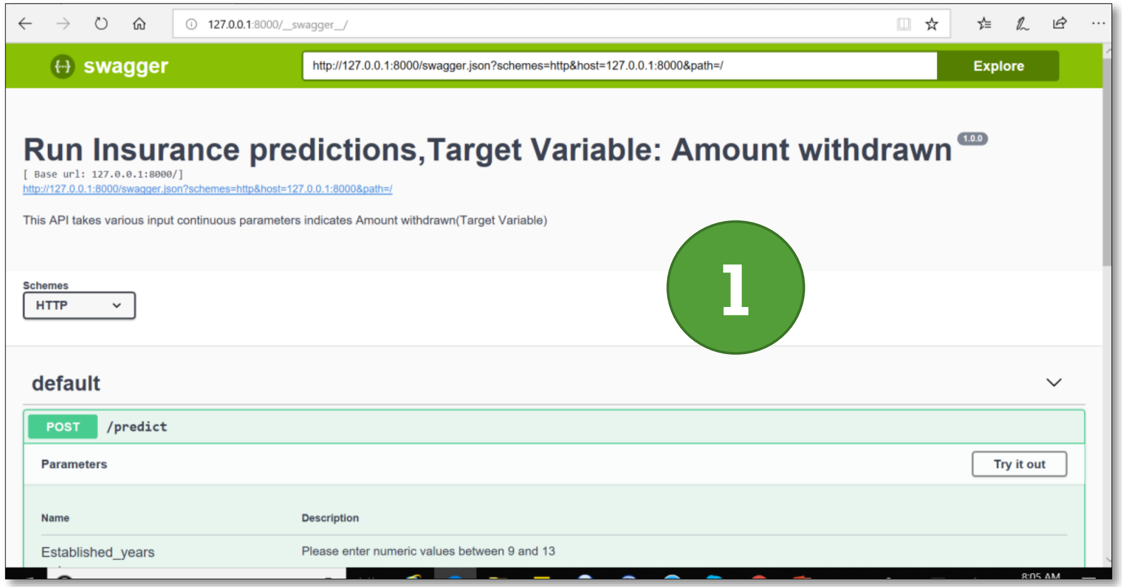
Admitted_Under	Age_Bracket	Established_years	(pr
Cardiology	30-40	11.00	
Cardiology	30-40	10.00	
Pulmonology	30-40	10.00	
Cardiology	30-40	13.00	
		10.00	
		11.00	
		10.00	

What do you want to do with
input_data_with_predictions,2019-07-13,.csv (1.3 MB)?
From: 127.0.0.1



	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	ium_	Illness_Ty	Hospital_L	Hostpital_	Hospital_E	No_Days_	Items_Not	Average_F	No_Times	Treatment	Hospital_Rating	Admitted_Under	Age_Brack	Establishe	(exp(prediction$predictions))	
	3574	Chronic	Urban	Multi Spec	Popular H	7	Histopath	864	3	Only Medi	4	Cardiology	30-40	11	213971.8485	
	2435	Chronic	Urban	Multi Spec	Popular H	6	Histopath	871	3	Only Medi	3	Cardiology	30-40	10	153632.81	
62	7740	Chronic	Urban	Multi Spec	Popular H	5	Physiother	977	4	Sugery	3	Pulmonology	30-40	10	128429.3012	
81	9718	Chronic	Town	Multi Spec	Popular H	8	Histopath	1079	3	Only Medi	4	Cardiology	30-40	13	219143.6822	
79	8042	Chronic	Town	Multi Spec	Popular H	6	Histopath	1240	2	Only Medi	4	Cardiology	50+	10	217422.1153	
61	9712	Chronic	Urban	Multi Spec	Popular H	7	Extra Atte	697	3	Only Medi	4	Pulmonology	30-40	11	163625.3221	
83	9729	Chronic	Urban	Multi Spec	Popular H	5	Histopath	1057	3	Only Medi	3	Cardiology	30-40	10	128399.2019	
78	14703	Chronic	Urban	Multi Spec	Popular H	5	Physiother	1382	2	Only Medi	3	Oncology	30-40	10	170882.056	
42	6899	Chronic	Town	Multi Spec	Popular H	7	Histopath	1392	2	Only Medi	4	Cardiology	50+	13	208287.7882	
72	9222	Chronic	Urban	Multi Spec	Popular H	5	Histopath	884	3	Only Medi	3	Pulmonology	30-40	13	134509.5196	
75	9369	Chronic	Urban	Single Spe	Popular H	6	Histopath	1091	3	Only Medi	3	Cardiology	30-40	12	139914.8019	
59	10718	Chronic	Urban	Multi Spec	Popular H	5	Histopath	812	2	Only Medi	3	Pulmonology	30-40	12	101110.7297	

Model Deployment – Plumber API



http://127.0.0.1:8000/swagger.json?schemes=http&host=127.0.0.1:8000&path=/

Run Insurance predictions,Target Variable: Amount withdrawn 1.0.0

[Base url: 127.0.0.1:8000/]
http://127.0.0.1:8000/swagger.json?schemes=http&host=127.0.0.1:8000&path=/

This API takes various input continuous parameters indicates Amount withdrawn(Target Variable)

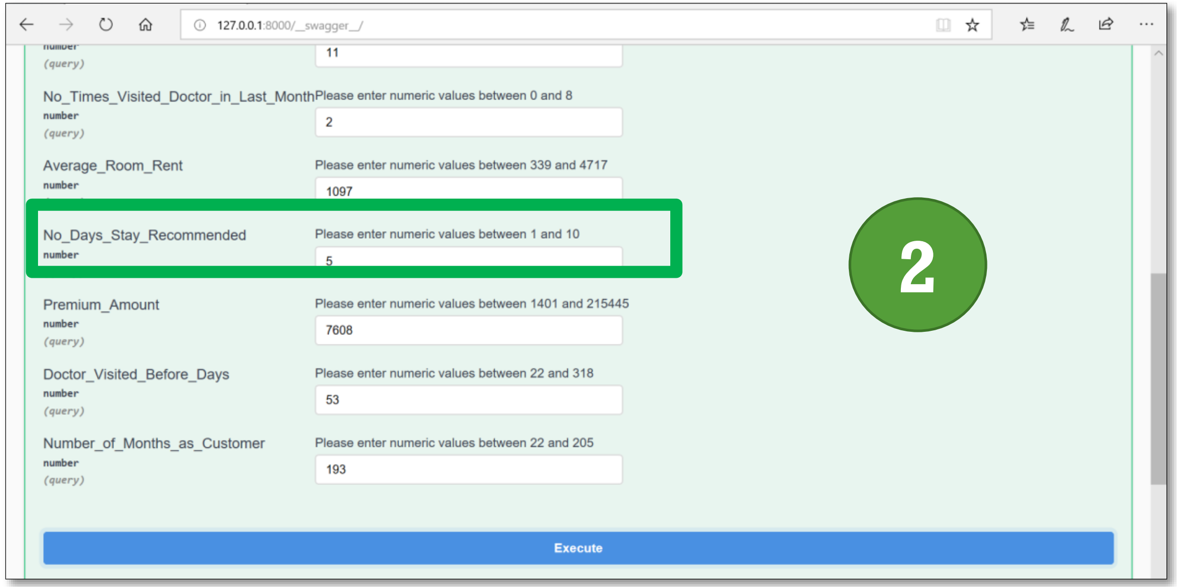
Schemes: HTTP

default

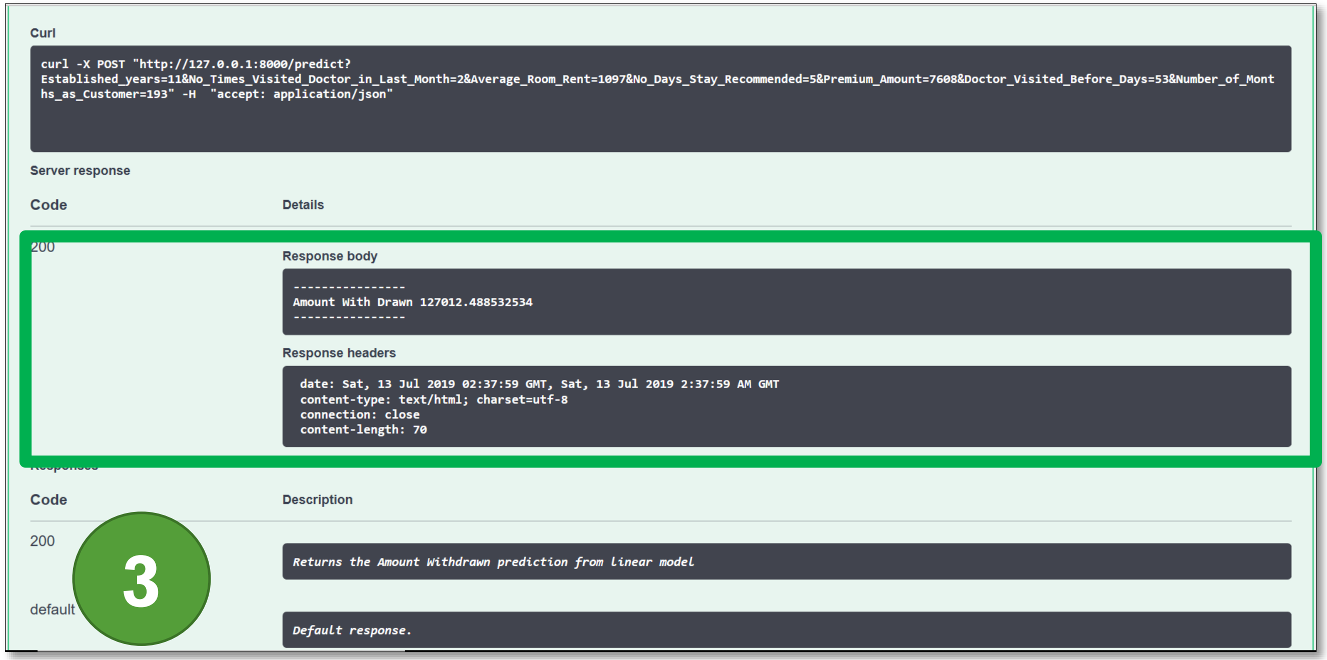
POST /predict

Parameters — Try it out

Name	Description
Established_years	Please enter numeric values between 9 and 13



number *(query)*	11
No_Times_Visited_Doctor_in_Last_Month number *(query)*	Please enter numeric values between 0 and 8 2
Average_Room_Rent number	Please enter numeric values between 339 and 4717 1097
No_Days_Stay_Recommended number	Please enter numeric values between 1 and 10 5
Premium_Amount number *(query)*	Please enter numeric values between 1401 and 215445 7608
Doctor_Visited_Before_Days number *(query)*	Please enter numeric values between 22 and 318 53
Number_of_Months_as_Customer number *(query)*	Please enter numeric values between 22 and 205 193



Curl

```
curl -X POST "http://127.0.0.1:8000/predict?
Established_years=11&No_Times_Visited_Doctor_in_Last_Month=2&Average_Room_Rent=1097&No_Days_Stay_Recommended=5&Premium_Amount=7608&Doctor_Visited_Before_Days=53&Number_of_Mont
hs_as_Customer=193" -H  "accept: application/json"
```

Server response

Code: 200

Details

Response body

```
---------------
Amount With Drawn 127012.488532534
---------------
```

Response headers

```
date: Sat, 13 Jul 2019 02:37:59 GMT, Sat, 13 Jul 2019 2:37:59 AM GMT
content-type: text/html; charset=utf-8
connection: close
content-length: 70
```

Responses

Code	Description
200	*Returns the Amount Withdrawn prediction from linear model*
default	*Default response.*

THANK YOU