

Proposal of class project

Predicting employee turnover

Mahsa Choopannezhad – student number: 98207477

Contents

Overview.....	2
Goals.....	2
Research questions	3
Resources.....	3
Introduction to dataset.....	4
Descriptive Analysis:	5
Predictive Analysis:.....	8
Decision tree	9
Cross validation on tree	10
Logistic regression	11
Random forest.....	12
Grid search.....	12
SVM	16
XG boost	17
Prescriptive Analysis:	19

Overview

پروژه درس مدل سازی و تصمیم گیری داده محور، تمرکز پروژه روی پیش بینی خروج کارمندان یک شرکت با توجه به دیتاهای موجود از کارمندان در دیتابیس HR است که به تصمیم گیری های واحد منابع انسانی برای اخذ اقدامات مناسب و در جهت دهی به استراتژی های منابع انسانی کمک به سزایی می کند. تمرکز تکنیکال پروژه روی رسیدن به یک مدل پیش بینی با نظارت با دقت مناسب است که بتوان به نتیجه آن برای اخذ استراتژی ها اعتماد کرد.

Goals

نرخ خروج کارمندان که تحت عنوان Employee turnover شناخته می شود یکی از متریک های مهمی است که در واحدهای منابع انسانی شرکت ها خیلی جدی مانیتور می شود و مدیران منابع انسانی همواره استراتژی های منابع انسانی سازمان را با هدف قرار دادن این متریک و پایین نگه داشتن آن تنظیم میکنند. علت هم این است که خروج کارمندان از شرکت و جایگزینی آنها یکی از مشکل های پرهزینه در سازمان هاست چرا که با از دست دادن یک کارمند، زمان زیادی صرف مصاحبه برای فرد جایگزین، در نظر گرفتن حقوق و مزایای احتمالا بالاتر می شود و همچنین چندماه از عملکرد تیم هم تا آماده سازی فرد جدید دچار افت بهره وری می شود.

در بعضی از کمپانی های بزرگ واحد منابع انسانی دارای یک تیم "HR Analytics" است که علاوه بر تمرکز روی موضوعات مختلف تحلیلی روی کم کردن نرخ خروج هم با استفاده از تحلیل دیتا کار می کنند. دیتاستی که برای پروژه انتخاب شده و در ادامه آن را معرفی می کنم از تیم HR analytics کمپانی IBM است که به صورت عمومی در اختیار بقیه قرار داده شده است.

هدف به طور کلی این است که با به کارگیری اتریبیوت های متفاوت موجود در دیتاست و احتمالا وارد کردن اتریبیوت های جدید بتوان به یک مدل پیشی بینی با دقت مناسب رسید که با دریافت اطلاعات جدید از یک کارمند **خروج یا عدم خروج** این کارمند در بازه زمانی کنونی را پیش بینی کند و البته این در کنار اطلاعات مفیدی در مرحله Descriptive است که میتوان در اختیار مدیران قرار داد.

Research questions

با بررسی و تحلیل این دتیاست به چه سوالاتی پاسخ می‌دهیم؟ تحلیل را در ۳ بخش اصلی انجام می‌دهیم:

Descriptive questions:

چه عوامل مشترکی در بین کارمندانی که سازمان (یا یک تیم خاص) را ترک میکنند وجود دارد؟

همبستگی بالا بین چه عواملی وجود دارد؟

Predictive questions:

با در دست داشتن اطلاعات ۳۴ متغیر دیگه آیا میتونیم پیش بینی کنیم که این فرد سازمان را ترک خواهد کرد یا نه و با چه دقتی؟

مهمترین عوامل تاثیرگذاری که در خروج افراد از سازمان نقش دارند چه عواملی هستند و میزان تاثیر گذرای اونها چقدر هست؟

Prescriptive questions:

چه سیاست ها و استراتژی های منابع انسانی براساس نتایج موجود میتوان اخذ کرد که به ننگه داشت کارمندان کمک کرد؟

Resources

<https://www.kaggle.com/code/faressayah/ibm-hr-analytics-employee-attribution-performance/notebook>

<https://github.com/mrc03/IBM-HR-Analytics-Employee-Attrition-Performance>

<https://www.kaggle.com/code/rtatman/machine-learning-with-xgboost-in-r/notebook>

<https://xgboost.readthedocs.io/en/stable/R-package/xgboostPresentation.html>

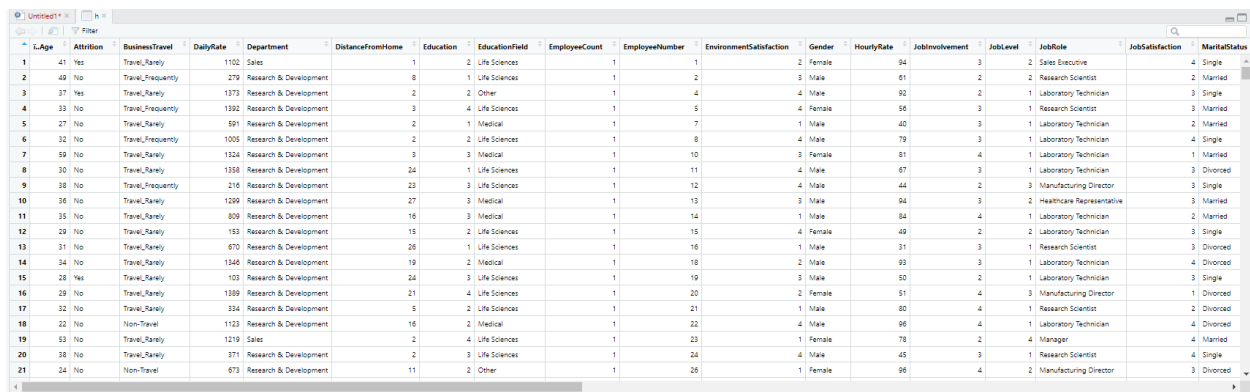
<https://www.geeksforgeeks.org/classifying-data-using-support-vector-machines-svms-in-r/>

<https://www.projectpro.io/recipes/use-svm-classifier-r>

Introduction to dataset

دیتاست شامل 1,470 رکورد و 35 ستون است که مربوط به اطلاعات کارمندان شرکت IBM است.

متغیر هدف، پیش بینی ستون Attrition است که Yes بودن آن به معنای خروج کارمند و No بودن آن به معنای عدم خروج کارمند است. ۳۴ ستون باقی مانده اطلاعاتی مثل سن، جنسیت، وضعیت تاهل، تحصیلات، دپارتمان، موقعیت های شغلی فرد، ارتقای شغلی فرد، نمره ارزیابی عملکرد فرد، حقوق و اطلاعاتی از این جنس از دیتابیس HR است.



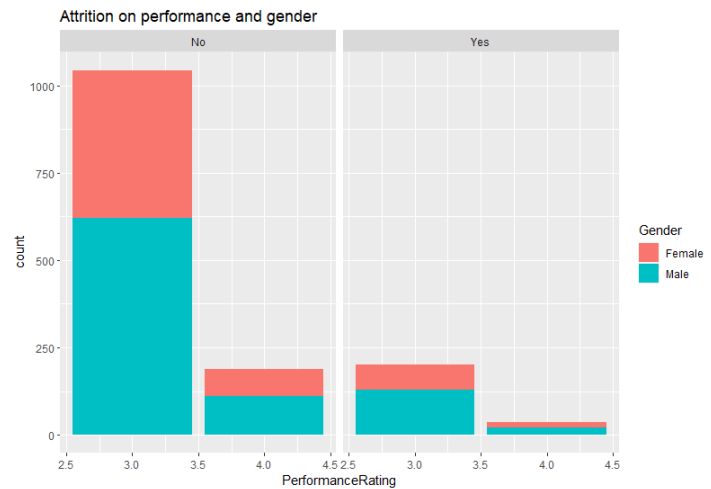
شکل- ۱- نمایی از دیتاست در Rstudio

با بررسی خلاصه دیتاست با دستور Summary و فاکتور کردن متغیرهای رشته ای، مشاهده می شود که:

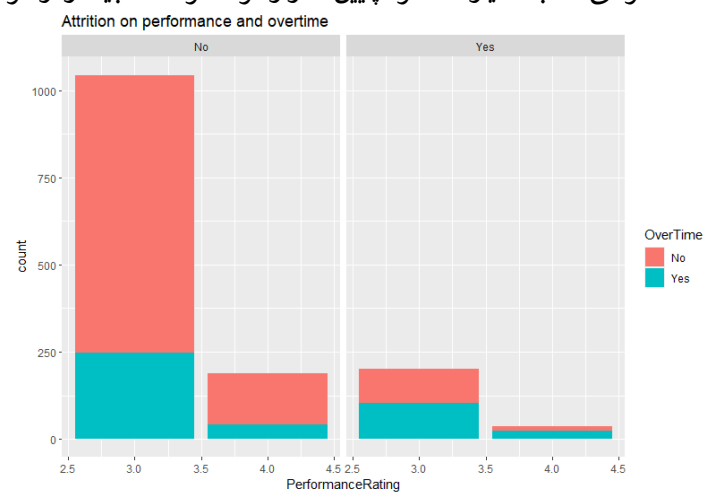
- دیتاست شامل هیچ رکورد ناقصی نیست اما از نظر تعداد داده روی دو سطح متفاوت از متغیر خروجی Attrition بالانس نیست.
- متغیر Over18 برای هر ۱۴۷۰ رکورد موجود مقدار ۲ را دارد و وضعیت کاملاً ثابتی برای همه دارد.
- متغیر EmployeeCount هم با میانگین و میانه و کمینه و بیشینه ثابت ۱ وضعیت ثابتی برای همه رکورد ها دارد.
- متغیر StandardHours هم با میانگین و میانه و کمینه و بیشینه ثابت ۸۰ وضعیت ثابتی برای همه رکورد ها دارد.
- متغیر EmployeeNumber با ۱۴۷۰ مقدار یکتا به عنوان شناسه هیچ اطلاعات اضافه ای برای تحلیل فراهم نمی کند.

برای ادامه روند پیش بینی باید در نظر داشت که این متغیرهای را حذف کرد.

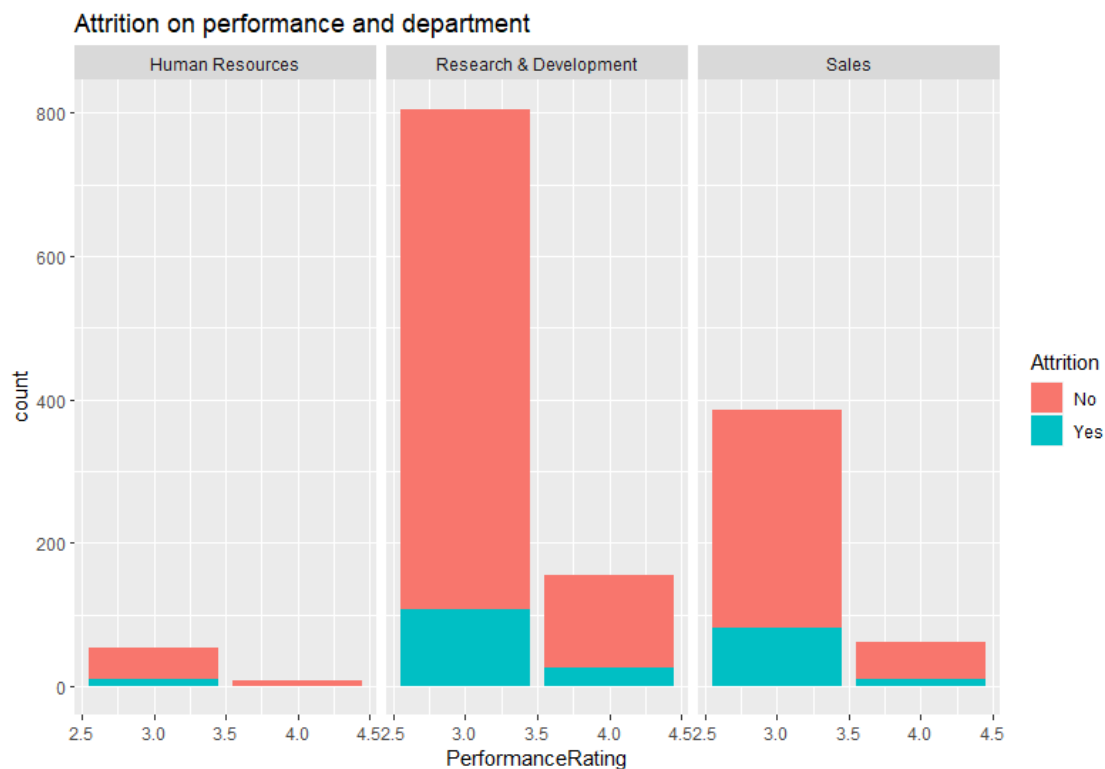
Descriptive Analysis:



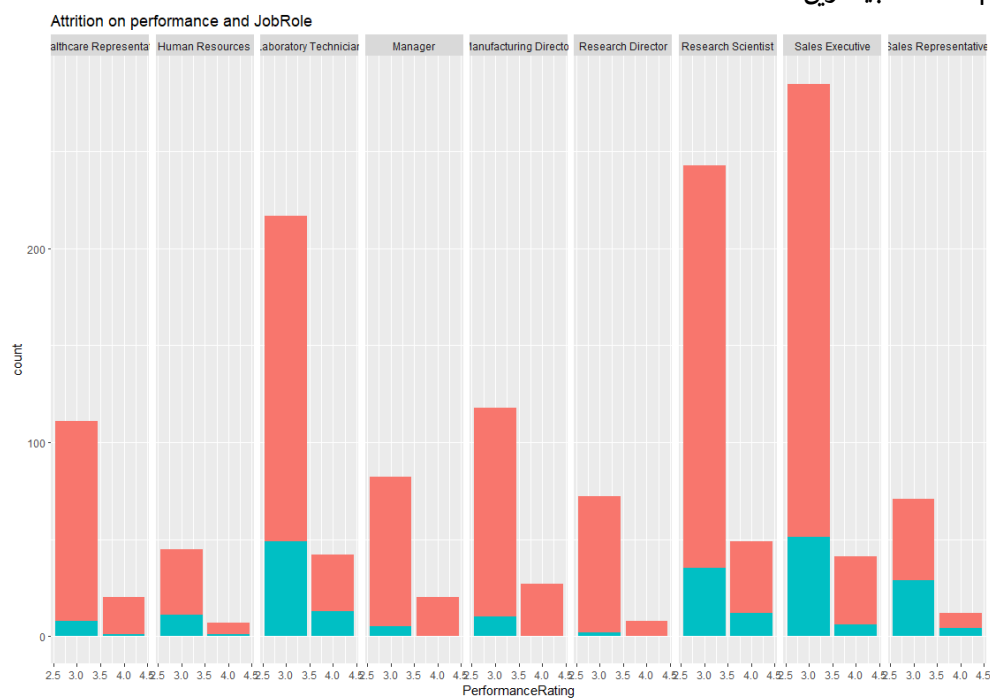
تعداد مردانی که با امتیاز عملکرد پایین، کار را ترک کرده اند بیشتر از مردانی است که نمره عملکرد بالایی داشته اند.



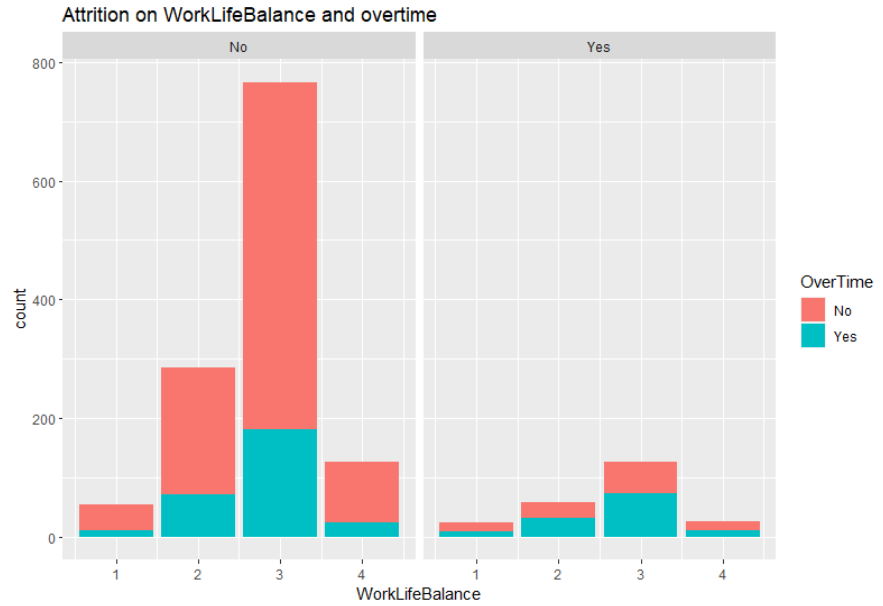
در بین کسانی که کار را ترک کرده اند، بیشتر افرادی که نمره عملکرد پایینی داشته اند، کمتر هم اضافه کاری کرده اند.



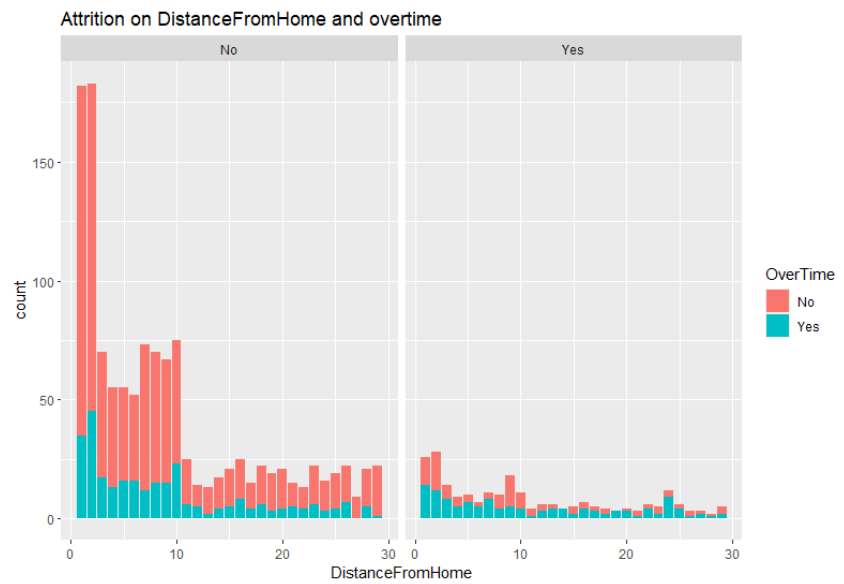
در بین دپارتمان های مختلف، تعداد افرادی که با نمره عملکرد بالا کار را ترک کرده اند در دپارتمان Research and development بیشترین است.



در بین پوزیشن های شغلی مختلف، laboratory technician و research scientist بیشترین ترک کار افراد با نمره عملکرد بالا را داشته اند.



ترند امتیاز تعادل کار و زندگی برای افرادی که ترک کار کرده اند و اضافه کاری داشته اند مشابه افرادی است که ترک کار نکرده داند.



تعداد افراد بیشتری که اضافه کاری کرده اند و فاصله کار تا خانه بیشتری داشته اند، از شرکت خارج شده اند. (به نسبت کسانی که از شرکت را ترک نکرده اند)

Predictive Analysis:

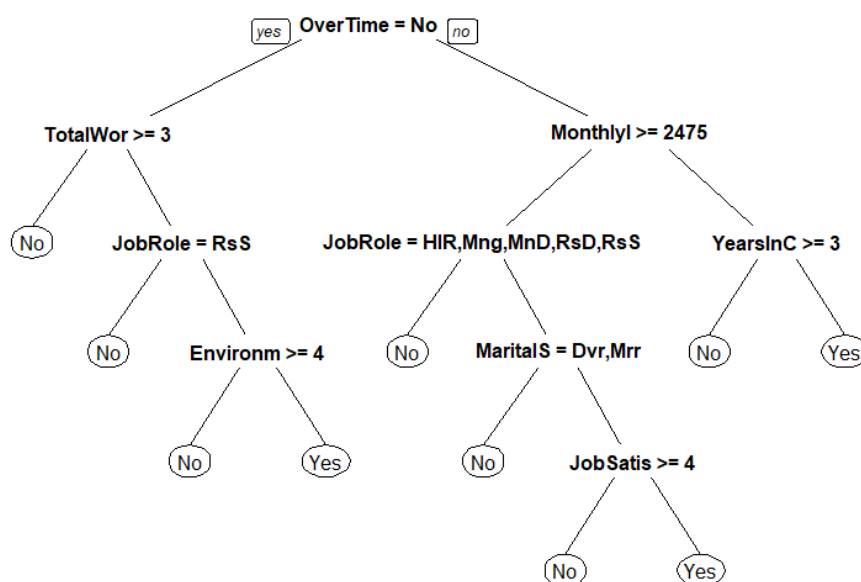
برای شروع تحلیل های مربوط به پیش بینی متغیر خروج افراد از شرکت ابتدا همبستگی بین متغیرهای متفاوت را با ماتریس همبستگی بررسی میکنیم. طبق این ماتریس:

- درآمد ماهانه با تعداد ساعت کارکردن کل وابستگی زیاد دارد.
- درآمد ماهانه با پوزیشن شغلی وابستگی زیاد دارد.
- تعداد سال هایی که فرد با مدیر فعلی کار کرده است با تعداد سال هایی که فرد در این سازمان بوده است و تعداد سال هایی که در این پوزیشن شغلی بوده است وابستگی زیاد دارد.
- سن با تعداد سال های کارکرد کل وابستگی زیاد دارد.
- انتخاب سهام با وضعیت مجرد بودن و پوزیشن شغلی وابستگی زیاد دارد.



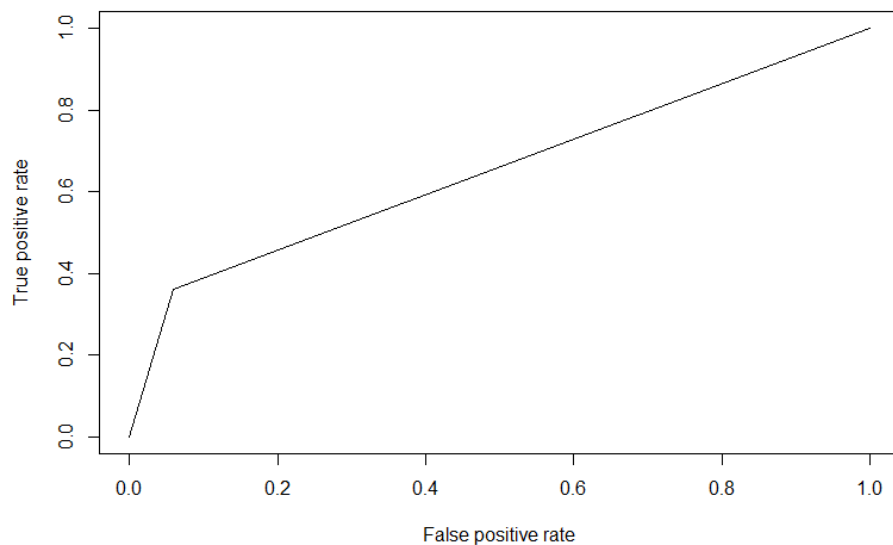
- متغیر خروجی Attrition وابستگی بالایی با متغیرهای اضافه کاری داشتن، تعداد سال هایی که فرد با مدیر فعلی بوده است، تعداد سال هایی که فرد در شرکت بوده است و تعداد سال هایی که فرد در پوزیشن فعلی خود بوده است.
- در ادامه با تعریف و تنظیم مدل های مختلف پیش بینی و با جداکردن ۸۵ درصد داده به عنوان دیتاست آموزش، سعی میشود که بهترین مدل انتخاب شود:

Decision tree



	<i>t_pred</i>	
	No	Yes
No	174	11
Yes	23	13

Accuracy	0.8462
Sensitivity	0.8832
Specificity	0.5417
Pos Pred Value	0.9405
Neg Pred Value	0.3611



باتوجه به خروجی مدل اولیه به نظر درخت تصمیم گیری مدل خوبی اصلا نباشد. در ادامه یک Cross validation هم انجام میشود که بتوان دید دقت مدل تغییر خواهد کرد یا نه:

Cross validation on tree

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was $cp = 0.0298$.

<i>pred_cv</i>		
	No	Yes
No	181	4
Yes	32	4

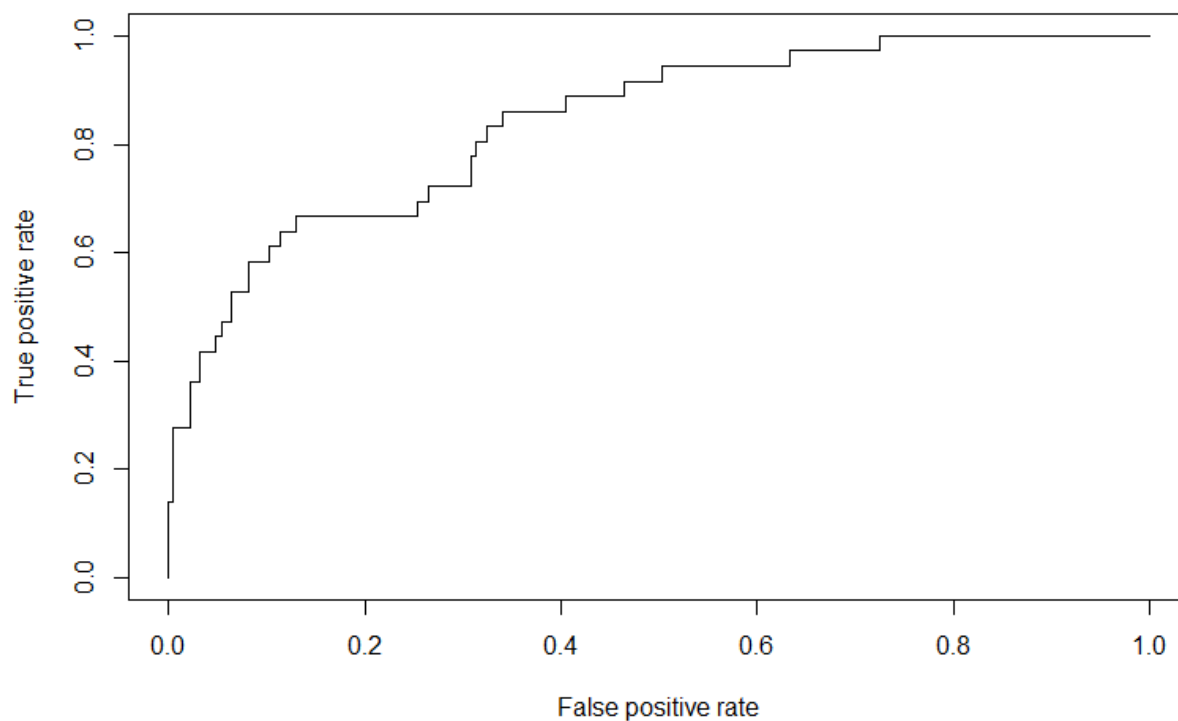
Auc: 0.5447447

با مقدار بهینه CP پیدا شده همچنان دقت مدل پایین است.

Logistic regression

با استفاده از مدل رگرسیون خطی، مدل را تشکیل می‌دهیم و با توجه به خروجی تاثیر گذار بودن متغیرهای مختلف سعی میشود به مدل بهینه رسید.

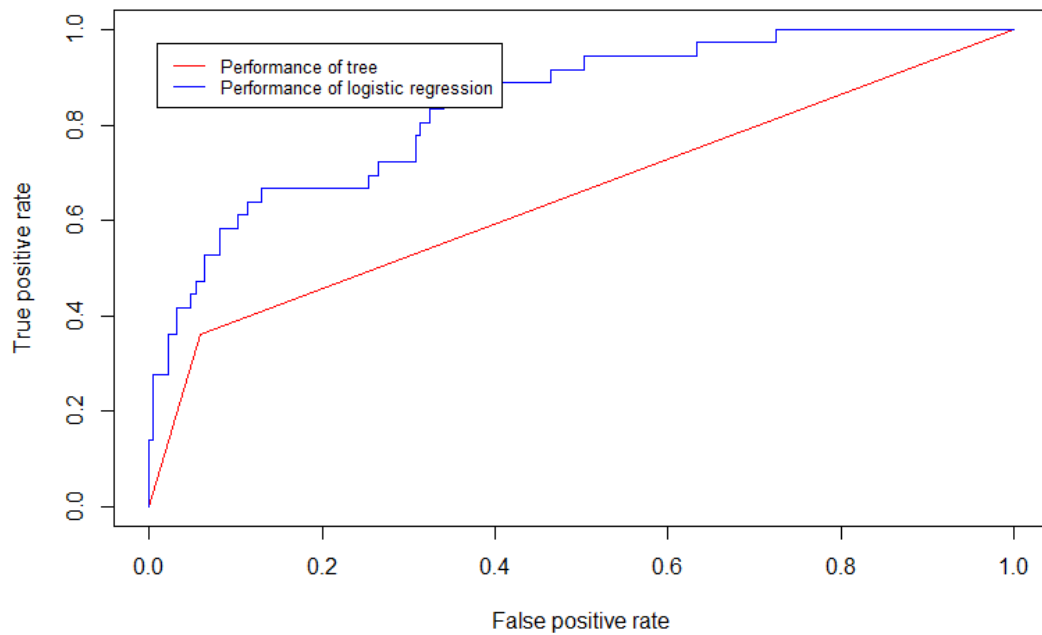
	<i>log_pred</i>	
	FALSE	TRUE
No	178	7
Yes	21	15



AUC: 0.8405405

مقایسه بین دو مدل درخت و رگرسیون خطی:

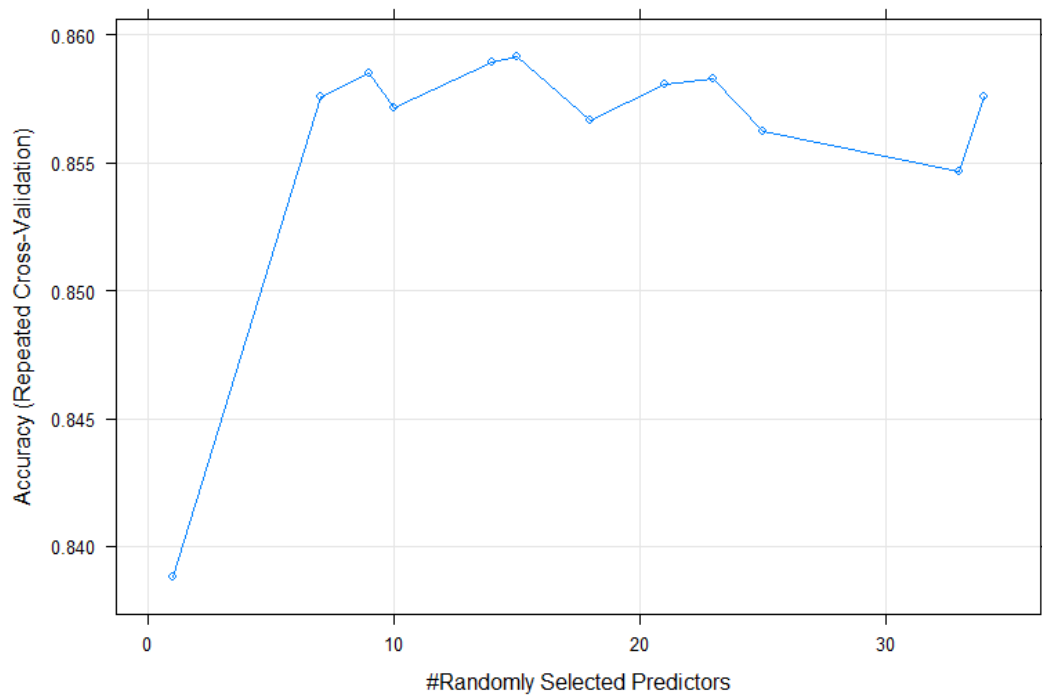
همانطور که مشاهده میشود مدل خطی مساحت زیر نمودار بسیار بزرگتر دارد و فعلا به عنوان مدل برگزیده انتخاب میشود. در ادامه چند مدل دیگر نیز امتحان و مقایسه میشوند.



Random forest

Grid search

به عنوان مدل بعدی برای پیدا کردن مقادیر بهینه برای مدل random forest از گرید سرچ استفاده میشود و با تست مقادیر مختلف برای mtry و با معیار قرار دادن دقت، نمودار دقت های تکرارهای متفاوت رسم میشود و در نتیجه mtry=15 برای مدل انتخاب میشود که دقت بالاتری دارد.



```
randomForest(formula = Attrition ~ ., data = train, nodsize = 5, mtry = 15, ntree = 50)
```

Type of random forest: classification

Number of trees: 500

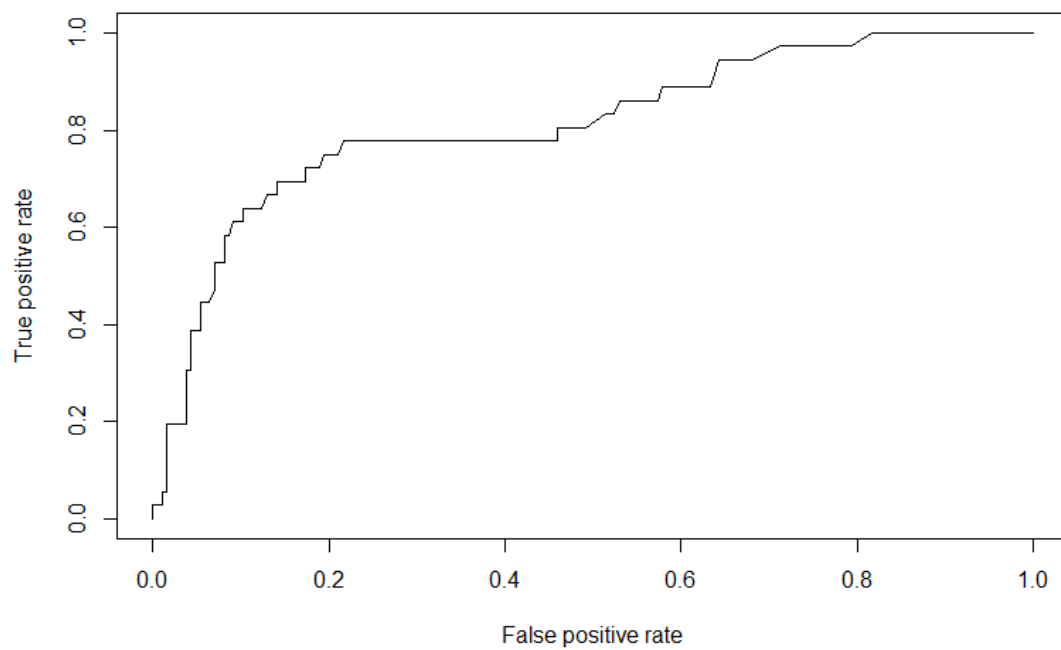
No. of variables tried at each split: 5

OOB estimate of error rate: 13.85%

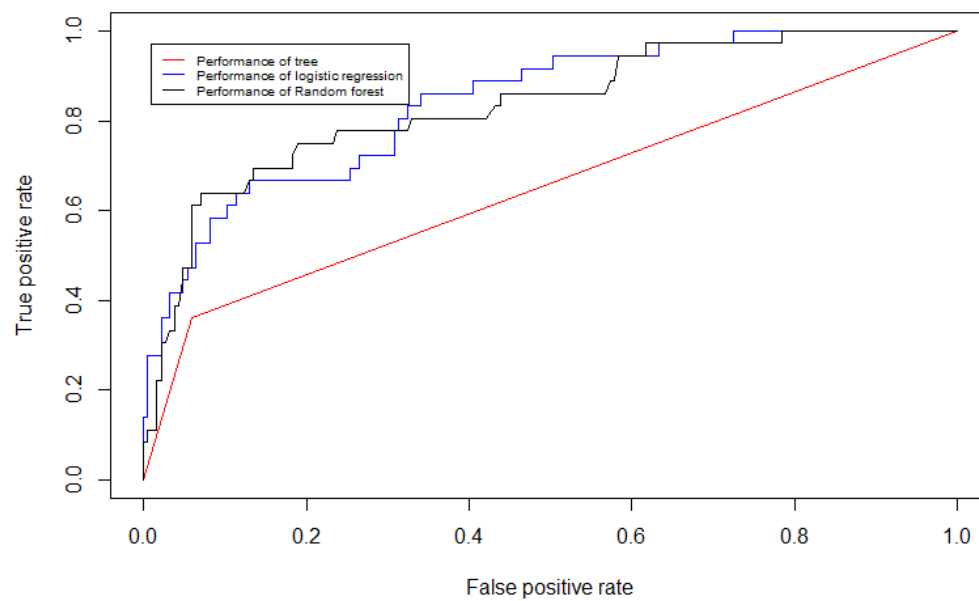
*Confusion
matrix:*

	No	Yes	class.error
No	1031	17	0.016221
Yes	156	45	0.776119

AUC: 0.862988



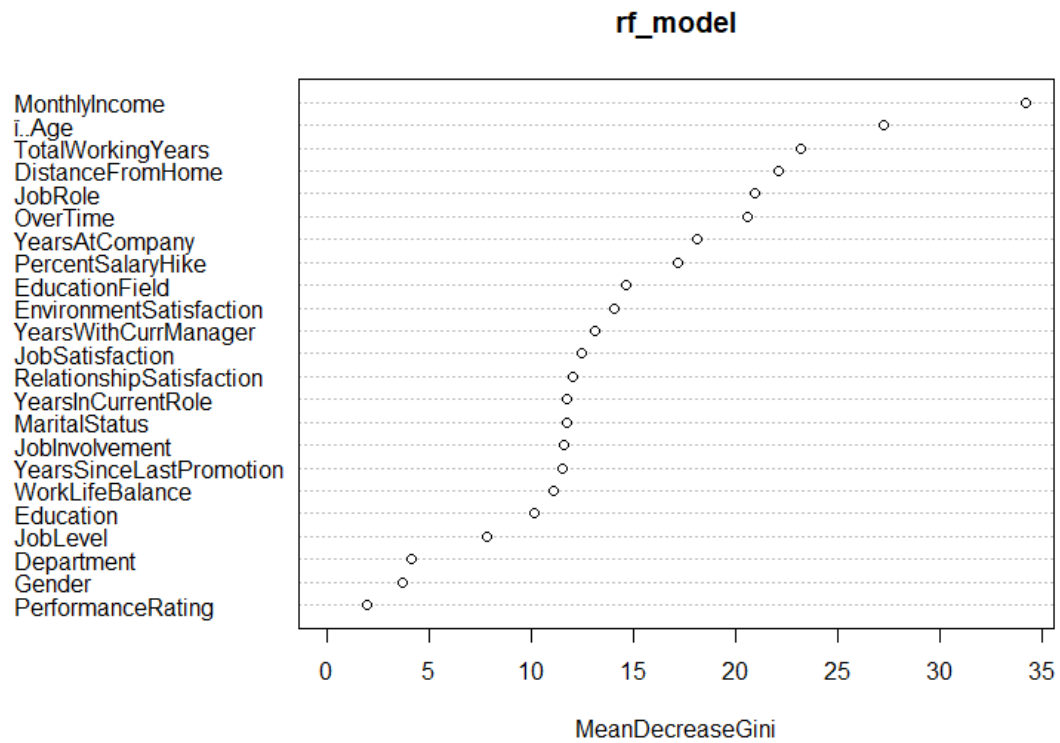
مقایسه بین ۳ مدل را انجام می‌دهیم:



در ادامه متغیرهایی که بیشترین اهمیت را در مدل random forest داشته اند را براساس اهمیت آنها مرتب میکنیم. امتیاز این ویژگی ها از این جهت حائز اهمیت است که برای تعریف اقدامات مناسب، میتوان از آنها استفاده کرد به طوری که تغییر در متغیرهایی که امتیاز بالایی دارند، نرخ خروج را که متغیر مورد پیش بینی ماست را از کلاسی به کلاسی دیگر سریع تر منتقل میکنند و این در اتخاذ تصمیمات و سیاست های راهبردی برای کاهش نرخ ترک کار، موثر خواهد بود. شکل هم درواقع میزان اثر گذاری هر متغیر را با امتیاز آن نشان میدهد.

مشاهده میشود که ۳ متغیر: درآمد ماهانه، سن، تعداد کل سال هایی که فرد کار کرده است، مهم ترین عواملی هستند که باعث میشوند نرخ ترک کار در شرکت تغییر کند و درواقع با تغییر آنها بزرگترین تاثیر را در پایین آمدن احتمال ترک کار یک فرد خواهیم داشت.

MeanDecreaseGini	
MonthlyIncome	34.22178
Age	27.28012
TotalWorkingYears	23.1829
DistanceFromHome	22.11347
JobRole	20.95723
OverTime	20.60739
YearsAtCompany	18.12048
PercentSalaryHike	17.18294
EducationField	14.64401
EnvironmentSatisfaction	14.07569
YearsWithCurrManager	13.15955
JobSatisfaction	12.45039
RelationshipSatisfaction	12.02338
YearsInCurrentRole	11.77096
MaritalStatus	11.74842
JobInvolvement	11.64305
YearsSinceLastPromotion	11.52773
WorkLifeBalance	11.12498
Education	10.12702
JobLevel	7.810295
Department	4.183543
Gender	3.695163
PerformanceRating	1.956617



SVM

مدل بعدی که مورد تحلیل قرار میگیرد مدل Super vector machine است. مشابه مدل قبل لازم است که مدل را تنظیم کنیم و برای همین از دوباره استفاده میکنیم.

```
svm(formula = Attrition ~ ., data = train, type = "C-classification", kernel = "linear")
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: linear

cost: 1

Number of Support Vectors: 409

<i>pred_svm</i>		
	No	Yes
No	181	4
Yes	24	12

Accuracy	0.8733
Sensitivity	0.8829
Specificity	0.75
Pos Pred Value	0.9784
Neg Pred Value	0.3333

XG boost

به عنوان مدل بعدی سراغ مدل گرادیان بوستد که با نام XG boost شناخته میشود رفتیم. با مشاهده خروجی میتوان دید که در round 14 کمترین RMSE اتفاق افتاده است.

```

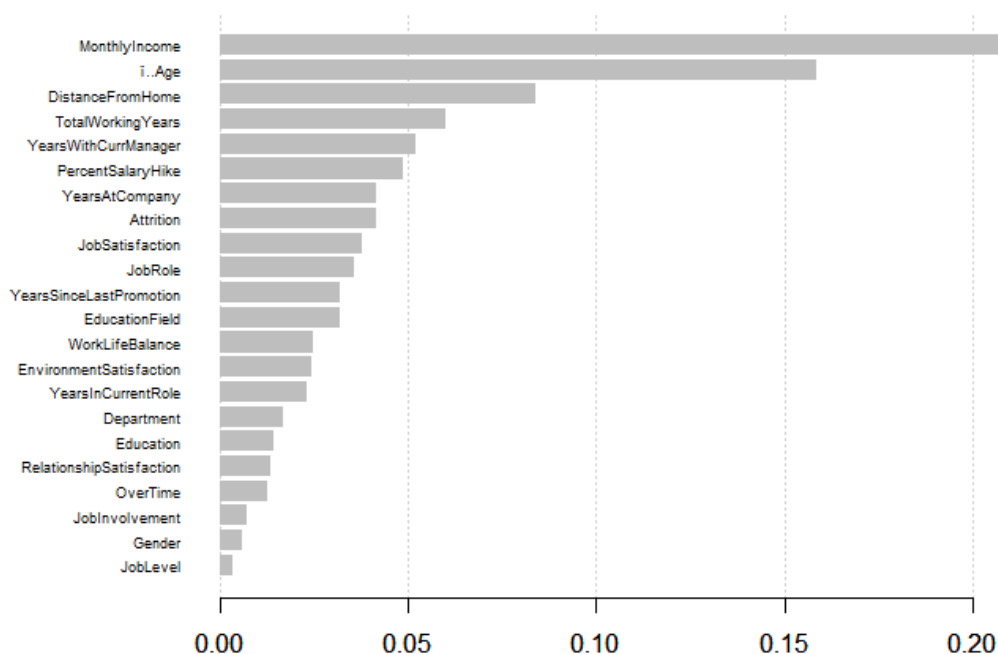
R 4.0.2 - ~/bin/R/nem6/Data driven decision making/projectdataset42/
> #fit xgboost model and display training and testing data at each round
> model = xgb.train(data = xgb_train, max_depth = 3, watchlist=watchlist, nrounds = 70)
[1] train-rmse:1.32425 test-rmse:1.373187
[2] train-rmse:1.052712 test-rmse:1.103909
[3] train-rmse:0.889163 test-rmse:0.940354
[4] train-rmse:0.793661 test-rmse:0.846602
[5] train-rmse:0.742208 test-rmse:0.793245
[6] train-rmse:0.712082 test-rmse:0.765381
[7] train-rmse:0.696558 test-rmse:0.749519
[8] train-rmse:0.687836 test-rmse:0.743904
[9] train-rmse:0.678225 test-rmse:0.739361
[10] train-rmse:0.674837 test-rmse:0.737844
[11] train-rmse:0.670803 test-rmse:0.737136
[12] train-rmse:0.668735 test-rmse:0.736307
[13] train-rmse:0.663931 test-rmse:0.733966
[14] train-rmse:0.657074 test-rmse:0.734178
[15] train-rmse:0.657632 test-rmse:0.734672
[16] train-rmse:0.654083 test-rmse:0.736862
[17] train-rmse:0.653280 test-rmse:0.736513
[18] train-rmse:0.647986 test-rmse:0.739570
[19] train-rmse:0.646894 test-rmse:0.740607
[20] train-rmse:0.643025 test-rmse:0.742272
[21] train-rmse:0.637402 test-rmse:0.740842
[22] train-rmse:0.633227 test-rmse:0.739999
[23] train-rmse:0.628469 test-rmse:0.742065
[24] train-rmse:0.626136 test-rmse:0.738694
[25] train-rmse:0.623719 test-rmse:0.738671
[26] train-rmse:0.620216 test-rmse:0.739297
[27] train-rmse:0.619489 test-rmse:0.739754
[28] train-rmse:0.617182 test-rmse:0.741867
[29] train-rmse:0.613382 test-rmse:0.741767
[30] train-rmse:0.612903 test-rmse:0.739842
[31] train-rmse:0.609891 test-rmse:0.740752
[32] train-rmse:0.607139 test-rmse:0.742000
[33] train-rmse:0.605990 test-rmse:0.742443
[34] train-rmse:0.603945 test-rmse:0.743153
[35] train-rmse:0.602956 test-rmse:0.744464
[36] train-rmse:0.599950 test-rmse:0.744553
[37] train-rmse:0.597903 test-rmse:0.746274
[38] train-rmse:0.594426 test-rmse:0.746288
[39] train-rmse:0.591768 test-rmse:0.749169
[40] train-rmse:0.588715 test-rmse:0.751596
[41] train-rmse:0.585512 test-rmse:0.753726
[42] train-rmse:0.584442 test-rmse:0.754129
[43] train-rmse:0.583694 test-rmse:0.753162
[44] train-rmse:0.580638 test-rmse:0.756925
[45] train-rmse:0.578186 test-rmse:0.757038
[46] train-rmse:0.575510 test-rmse:0.757892
[47] train-rmse:0.573364 test-rmse:0.759211
[48] train-rmse:0.571011 test-rmse:0.759680
[49] train-rmse:0.569380 test-rmse:0.761279
[50] train-rmse:0.568596 test-rmse:0.762617
[51] train-rmse:0.566392 test-rmse:0.763873
[52] train-rmse:0.563755 test-rmse:0.763061
[53] train-rmse:0.561413 test-rmse:0.764989
[54] train-rmse:0.559247 test-rmse:0.766119
[55] train-rmse:0.557894 test-rmse:0.765711
[56] train-rmse:0.554621 test-rmse:0.763973

```

با استخراج مهمترین ویژگی هایی که بر متغیر مورد پیش بینی ما یعنی Attrition تاثیر گذار هستند، میتوان به مهم ترین کاندید برای تغییر در تصمیم گیری ها رسید. شکل نیز تاثیر هریک را برحسب Gain نشان میدهد.

مشاهده میشود که ۳ متغیر: درآمد ماهانه، سن، فاصله کار از خانه مهم ترین عواملی هستند که باعث میشوند نرخ ترک کار در شرکت تغییر کند و درواقع با تغییر آنها بزرگترین تاثیر را در پایین آمدن احتمال ترک کار یک فرد خواهیم داشت.

<i>Feature</i>	<i>Gain</i>	<i>Cover</i>	<i>Frequency</i>
<i>MonthlyIncome</i>	0.231972	0.23134	0.206897
<i>i..Age</i>	0.158514	0.130525	0.12069
<i>DistanceFromHome</i>	0.083846	0.050893	0.088362
<i>TotalWorkingYears</i>	0.059983	0.102102	0.077586
<i>YearsWithCurrManager</i>	0.052026	0.074802	0.05819
<i>PercentSalaryHike</i>	0.04875	0.036525	0.05819
<i>YearsAtCompany</i>	0.041431	0.069548	0.049569
<i>JobSatisfaction</i>	0.037643	0.017984	0.030172
<i>JobRole</i>	0.035829	0.03481	0.043103
<i>YearsSinceLastPromotion</i>	0.031783	0.029392	0.030172
<i>EducationField</i>	0.031753	0.012485	0.030172
<i>WorkLifeBalance</i>	0.024931	0.005498	0.021552
<i>EnvironmentSatisfaction</i>	0.024125	0.012799	0.030172
<i>YearsInCurrentRole</i>	0.022958	0.0599	0.038793
<i>Department</i>	0.01672	0.008495	0.015086
<i>Education</i>	0.014398	0.005223	0.010776
<i>RelationshipSatisfaction</i>	0.013328	0.016728	0.012931
<i>OverTime</i>	0.012477	0.028457	0.019397
<i>JobInvolvement</i>	0.007063	0.014845	0.015086
<i>Gender</i>	0.0057	0.010687	0.008621
<i>JobLevel</i>	0.003406	0.007025	0.008621



Prescriptive Analysis:

با بدست آوردن مدل های مختلف پیش بینی و ارزیابی و مقایسه آنها، دو مدلی که میتوان پیش نهاد کرد که براساس آنها احتمال خروج هر فرد باتوجه به داشتن بقیه متغیرها، پیش بینی شود؛ مدل random forest و XG boost است.

براساس این مدل ها وقتی داده های جدید از فرد را به مدل مورد تخمین بدهیم، مدل با دقت حدود ۸۶ درصد قادر خواهد بود که احتمال خروج فرد را با وضعیت کنونی متغیرهای داده شده پیش بینی کند و بعد در ادامه میتوان افرادی را که احتمال خروج بالا برای آنها پیش بینی شده است را مدنظر قرار داد و با تغییر ۳ یا ۵ متغیری که در لیست متغیرهای تاثیر گذار مدل هستند، احتمال خروج آنها را کاهش داد و این باعث صرفه جویی مالی قابل ملاحظه در شرکت خواهد شد.

اما به عنوان پیشنهاد های کلان تر میتوان موارد زیر را هم باتوجه به لیست تاثیر گذاری متغیرها در نظر گرفت:

- باتوجه به رتبه اول بودن درآمد ماهانه در لیست متغیرهای تاثیر گذار بهتر است که شرکت نسبت به مرور و بازبینی رنج حقوق های خود اقدام کند و متناسب با بودجه موجود و وضعیت بقیه شرکت های رقیب درآمد ماهانه افراد را تنظیم مجدد کند تا از خروج و پیوستن آنها به شرکت های رقیب جلوگیری کند.

-
- رتبه دوم در عوامل تاثیرگذار را سن دارد که باتوجه به خروج بیشتر افراد با سن ۲۰-۳۰ بهتر است که شرکت برنامه های بلند مدت برای رشد و توسعه نیروهای جوان خود داشته باشد تا نرخ خروج آنها را کاهش دهد.
 - مورد بعدی برای ایجاد تغییر میتواند متغیر فاصله خانه تا کار باشد و برای کاهش نرخ خروج افرادی که فاصله خانه تا محل کار آنها زیاد است میتوان کمک هزینه حمل و نقل و یا سرویس قرار داد.
- در نهایت یک برنامه بلند مدت با تعریف اقدامات مناسب باتوجه به امتیازهای بدست آمده برای لیست عوامل تاثیرگذار مدل مورد نظر میتوان تدوین کرد و در اختیار مدیر منابع انسانی شرکت قرار داد.