HR-аналитика: увольнение сотрудников

Набор данных о 1470 сотрудниках: 237 уволившихся и 1233 еще работающих.

Кроме того, в наборе данных 34 показателях, таких как: возраст, удовлетворение от работы, уровень зарплаты, баланс работы и личной жизни и т.д. Задача анализа – выяснить причины увольнения сотрудников и научится прогнозировать уход сотрудников.

(Это вымышленный набор данных, созданный специалистами по данным IBM. Данные открытые. Ссылка на датасет: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset )

1. При первоначальном анализе данных с помощью запросов SQL к базе данных видно взаимосвязь увольнения сотрудников со следующими факторами:

* **возраст** (больше всего уволившихся сотрудников в возрасте 18-21 (примерно 50% от всех сотрудников такого возраста). Далее процент уволившихся уменьшается до 30% и после 33 лет резко падает),
* **зарплата** (чаще всего увольняются молодые сотрудники (до 33 лет) и сотрудники с невысокой зарплатой (до 3000),
* **должность** (большинство уволившихся были на должности Sales Representative),
* **переработки** (среди уволившихся больше тех, кто перерабатывал),
* **уровень работы** (уволившихся сотрудников больше всего тех, кто относится к 1-ому уровню работы (к сожалению, нет расшифровки что бы это могло значить), на втором уровне работы работает большинство сотрудников),
* **сколько лет проработал в компании** (большая доля сотрудников (по отношения к работающим) увольняется в первый и второй год работы в компании (36% и 35%)).

Из этих взаимосвязей можно предположить, следующее:

* Адаптационных мероприятий при устройстве на работу в компанию и при переходе на новую должность недостаточно для удержания специалистов, возможно их нужно доработать.
* Сотрудникам не хватает возможностей для карьерного развития внутри компании. Необходимо разработать карьерную карту.
* Возможно, руководителям не хватает лидерских компетенций. Стоит пристальнее присмотреться к их стилю руководства и возможно провести дополнительное обучение
* Имеет место выгорание сотрудников. Необходимо понять почему сотрудникам приходится перерабатывать и разработать мероприятия по решению проблем с выгоранием

2. Из расчета прогноза увольнения сотрудников в BigQuery и в Python можно увидеть, какие признаки разные модели машинного обучения в разных системах посчитали значимыми:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | BQ Логистич регрессия 2 | BQ Логистич регрессия 3 | BQ XGBOOST 1 gain | BQ XGBOOST 2 gain | Python Логистическая регрессия | Python LDA | Python XGBOOST | Python Gradient Boosting | Частота |
| OverTime | 0,3091 | 0,2725 | 0,1976 | 0,2076 | 0,3355 | 0,0819 | 0,1536 | 0,3289 | 8 |
| JobLevel |  |  | 0,3979 | 0,3468 | 0,1573 |  | 0,4232 |  | 4 |
| TotalWorkingYears |  |  | 0,1265 | 0,1582 |  | 0,2028 |  | 0,1694 | 4 |
| JobRole | 0,2353 | 0,2165 |  |  |  |  | 0,1469 |  | 3 |
| DistanceFromHome | 0,0798 | 0,1013 |  |  |  | 0,2242 |  |  | 3 |
| Age | 0,0699 |  |  |  |  | 0,3381 |  | 0,1544 | 3 |
| MaritalStatus | 0,2169 | 0,1911 |  |  |  |  |  |  | 2 |
| StockOptionLevel |  |  | 0,1345 | 0,1721 |  |  |  |  | 2 |
| NumCompaniesWorked | 0,0890 | 0,1242 |  |  |  |  |  |  | 2 |
| JobInvolvement |  |  |  |  | 0,1581 |  | 0,1369 |  | 2 |
| YearsAtCompany |  |  |  |  |  | 0,1530 |  | 0,1111 | 2 |
| Gender |  |  | 0,1435 |  |  |  |  |  | 1 |
| Department |  |  |  | 0,1154 |  |  |  |  | 1 |
| YearsSinceLastPromotion |  | 0,0944 |  |  |  |  |  |  | 1 |
| BusinessTravel |  |  |  |  | 0,1754 |  |  |  | 1 |
| YearsWithCurrManager |  |  |  |  | 0,1737 |  |  |  | 1 |
| Department\_Sales |  |  |  |  |  |  | 0,1394 |  | 1 |
| MonthlyIncome |  |  |  |  |  |  |  | 0,2363 | 1 |

Чтобы выбрать на данные какой модели ориентироваться, рассмотрим метрики моделей:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Recall** | **Precision** | **Accuracy** | **Roc\_auc** |
| BQ Логистич регрессия 1 | 0,2889 | 0,8667 | 0,8555 | 0,8296 |
| BQ Логистич регрессия 2 | 0,1273 | 1 | 0,80166 | 0,8 |
| BQ Логистич регрессия 3 | 0,6618 | 0,5 | 0,7785 | 0,8081 |
| BQ XGBOOST 1 | 0,25 | 0,5 | 0,8889 | 0,7713 |
| BQ XGBOOST 2 | 0,5 | 0,4 | 0,8611 | 0,7633 |
| Python Логистическая регрессия | 0,4219 | 0,7297 | 0,8669 | 0,6936 |
| Python Логистическая регрессия тюн | 0,3594 | 0,7931 | 0,8669 | 0,6693 |
| Python Логистическая регрессия финал | 0,5312 | 0,9444 | 0,9093 | 0,7622 |
| Python Градиентный бустинг | 0,2656 | 0,68 | 0,8442 | 0,619 |
| Python Градиентный бустинг тюн | 0,0312 | 0,6667 | 0,8215 | 0,5139 |
| Python Градиентный бустинг финал | 0,1875 | 1 | 0,8527 | 0,5938 |
| Python XGBOOST | 0,3438 | 0,7857 | 0,864 | 0,6615 |
| Python XGBOOST tuned | 0,5156 | 0,6471 | 0,8612 | 0,7267 |
| Python XGBOOST final | 0,75 | 0,7619 | 0,9122 | 0,849 |
| Python Linear Discriminant Analysis | 0,4375 | 0,5957 | 0,8442 | 0,6859 |
| Python Linear Discriminant Analysis tuned | 0 | 0 | 0,8187 | 0,5 |
| Python Linear Discriminant Analysis final | 0 | 0 | 0,8187 | 0,5 |
| Python Blended model | 0,3438 | 0,6875 | 0,8527 | 0,6546 |
| Python Blended model tuned | 0,3438 | 0,6875 | 0,8527 | 0,6546 |
| Python Blended model final | 0,7031 | 0,9375 | 0,9377 | 0,8464 |
| **Метрики моделей на новых данных** | | | | |
|  | **Recall** | **Precision** | **Accuracy** | **Roc\_auc** |
| Python Blended model final on unseen data | 0,5111 | 0,7667 | 0,9014 | 0,7415 |
| Python Gradient Boosting Classifier final on unseen data | 0,1333 | 1 | 0,8673 | 0,83 |
| Python XGBOOST final on unseen data | 0,5556 | 0,5618 | 0,8673 | 0,8441 |
| Python Linear Discriminant Analysis final on unseen data | 0 | 0 | 0,8469 | 0,6967 |
| Python Logistic Regression final on unseen data | 0,4444 | 0,8333 | 0,9014 | 0,8591 |

В нашем случае наиболее важной является метрика Recall – она показывает, насколько часто модель предсказывает положительные случаи (в нашем случае случаи увольнения). Если сравнивать модели по этой метрике, то:

1. Логистическая регрессия в BigQuery делает прогноз лучше, чем логистическая регрессия в Python PyCaret.
2. А XGBOOST наоборот делает прогноз лучше в Python PyCaret, чем в BigQuery.
3. Кроме того, по метрике Recall XGBOOST в Python PyCaret делает прогноз лучше чем любая другая модель в Python PyCaret, даже лучше чем та, которая используют в себе смесь моделей (Blended model построена на основе Naive Bayes, Ridge Classifier, CatBoost Classifier, Linear Discriminant Analysis, Logistic Regression).

Значит имеет смысл рассматривать значимые параметры модели XGBOOST в Python PyCaret:

JobLevel\_1 – Уровень работы «1»

JobLevel\_3 – Уровень работы «3»

OverTime\_True – Переработки «были»

JobRole\_Sales Representative – Должность «Торговый представитель»

Department\_Sales – Департамент «Продаж»

JobInvolvement\_2 – Вовлеченность в работу «Средняя»

Эти данные в целом подтверждают открытые закономерности в данных, что увольнение сотрудника зависит от:

* Переработок
* Уровня работы
* Должности

Что мы узнали новое из тех признаков, которые модель машинного обучения посчитала важными:

1. Сотрудники увольняются не только с уровня работы 1, но и с уровня работы 3.
2. Департамент
3. Уровень вовлеченности в работу «Средняя». Этот фактор требует дополнительного исследования, потому что всё-таки среди уволившихся больше сотрудников с уровнем вовлеченности 3, а соотношение уволившихся к работающим больше для уровня вовлеченности 1.