Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _____

КАФЕДРА _____

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Предсказание возможности решения работника сменить место работы

Студент ИУ5-65Б _____ **Домрачева К.Г.**
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ **Гапанюк Ю.Е.**
(Подпись, дата) (И.О.Фамилия)

Консультант _____ **Гапанюк Ю.Е.**
(Подпись, дата) (И.О.Фамилия)

2023 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Заведующий кафедрой _____
(Индекс)
_____ _____
(И.О.Фамилия)
« _____ » _____ 20 ____ г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме ____Предсказание возможности решения работника сменить место работы_____

_____

_____

Студент группы __ИУ5-65Б_____

_____Домрачева Ксения Григорьевна_____
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)
_____Исследоваткльская_____
Источник тематики (кафедра, предприятие, НИР) _____НИР_____

График выполнения НИР: 25% к ___ нед., 50% к ___ нед., 75% к __ нед., 100% к ___ нед.

Техническое задание _____
_____Исследовать методы машинного обучения для решения задачи классификации_____
_____
_____
_____

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на __32__ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)
_____
_____
_____
_____

Дата выдачи задания « 13 » февраля 2023 г.

Руководитель НИР _____ ____Гапанюк Ю.Е._____
(Подпись, дата) (И.О.Фамилия)

Студент _____ ____**Домрачева К.Г.**_____
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Содержание

Введение

Кадровые движения сотрудников в современном мире сильно волнуют работодателей. Изучив данные о движении сотрудников, можно попробовать ответить на крайне важные для работодателя вопросы:

1. Как удержать сотрудников?
2. Как определить структуру заработной платы?
3. Как определить структуру отпусков?

Исследователи, аналитики данных и специалисты по персоналу могут получить ценную информацию по данным, предоставленным кадровыми специалистами.

В данной работе я буду использовать обезличенные данные, собранные в трех городах специалистами отдела кадров. Я выделила один из вопросов кадрового движения целью своей работы: построить модель машинного обучения, которая сможет предсказывать решение сотрудника сменить место работы. Я буду использовать алгоритмы классификации для определения факторов риска смены работы, включая образование, пол, возраст, город, уровень оплаты и опыт работы.

Результатом данной работы станет эффективная модель, которая может помочь оценить риск принятия сотрудником решения сменить работу.

Для достижения поставленной цели были определены следующие этапы:

1. Поиск и выбор набора данных для построения моделей машинного обучения для решения задачи регрессии или классификации.
2. Проведение разведочного анализа данных.
3. Выбор признаков, подходящих для построения моделей.
4. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
5. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.
6. Выбор метрик для последующей оценки качества моделей.

7. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии.
8. Формирование обучающей и тестовой выборок на основе исходного набора данных.
9. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров и оценка качества моделей на основе тестовой выборки.
10. Подбор гиперпараметров для выбранных моделей. Построение оптимальных моделей.
11. Формирование выводов о качестве построенных моделей на основе выбранных метрик.

Постановка задачи

Данная работа по машинному обучению направлена на решение задачи классификации, а именно, предсказание риска принятия сотрудником решения о смене места работы.

Я взяла за основу данные о работниках, которые приняли решение остаться на прежнем месте или сменить текущее место работы. Данные включают информацию о таких факторах, как образование, пол, возраст, город, уровень оплаты и опыт работы. Каждый сотрудник может быть классифицирован как потенциально рискующий сменить работу и наоборот.

Целью задачи является создание модели машинного обучения, которая будет использовать имеющиеся данные для предсказания риска принятия данного решения. Для этого мы будем использовать различные алгоритмы классификации, такие как K ближайших соседей, метод опорных векторов, дерево решений, случайный лес и градиентный бустинг. Модель должна обучаться на тренировочных данных и проверяться на тестовых данных для оценки ее точности и эффективности.

Результатом работы должна быть модель, которая сможет предсказывать возникновения решения сменить место работы с высокой точностью и помочь работодателям принимать меры для предотвращения таких ситуаций.

Выполнение работы

Для решения задачи классификации был выбран набор данных, содержащий информацию о сотрудниках.

В наборе данных присутствуют следующие столбцы:

1. Education: образовательная квалификация сотрудников;
2. Joining Year: год, когда каждый сотрудник присоединился к компании, с указанием стажа работы.
3. City: место или город, где находится или работает каждый сотрудник.
4. Payment Tier: категоризация сотрудников по разным уровням заработной платы.
5. Age: возраст каждого сотрудника, предоставляющий демографическую информацию.
6. Gender: пол.
7. Ever Benched: указывает, находился ли сотрудник когда-либо временно безработным.
8. Experience in Current Domain: количество лет опыта сотрудников в текущей области.
9. Leave or Not: целевое значение, определяющее, принял ли сотрудник решение поменять место работы.

Данный датасет использован для решения задачи классификации.

Загружаем данные, получаем обую информацию о датасете и делаем предположения о влиянии признаков на целевую переменную. В наборе данных содержится 4653 строк и 9 столбцов.

- Пропусков в данных нет;
- Дублирующиеся строки удалим. После удаления осталось 2764 строк.

Строим график pairplot для визуализации распределения данных попарно для множества колонок.

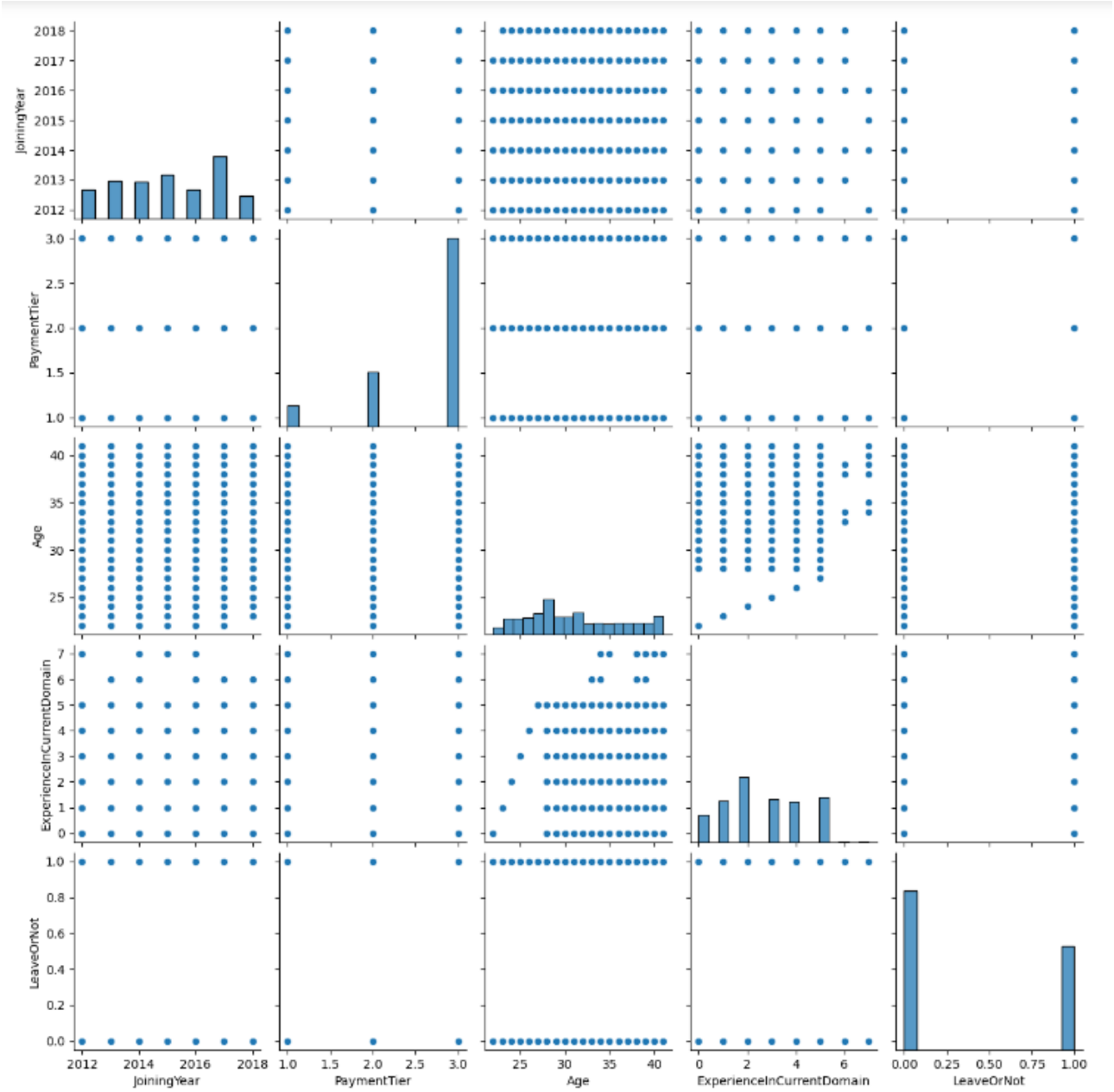


Рисунок 1 - Визуализация распределения данных попарно для множества колонок

Проверяем сбалансированы ли классы в нашем наборе данных. Получаем следующую гистограмму:

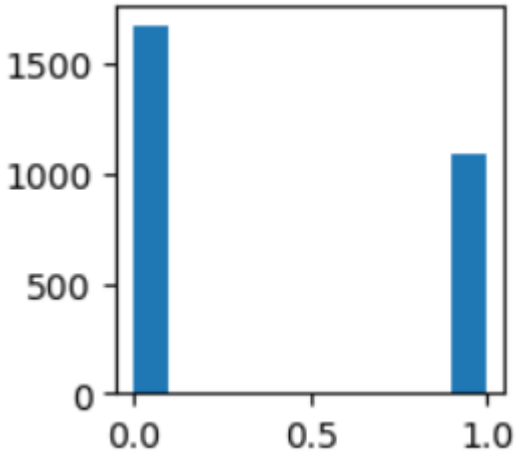


Рисунок 2 - Гистограмма классов

Видим, что классы достаточно сбалансированы.

Проведем исследование данных на основе парных гистограмм по категориям.

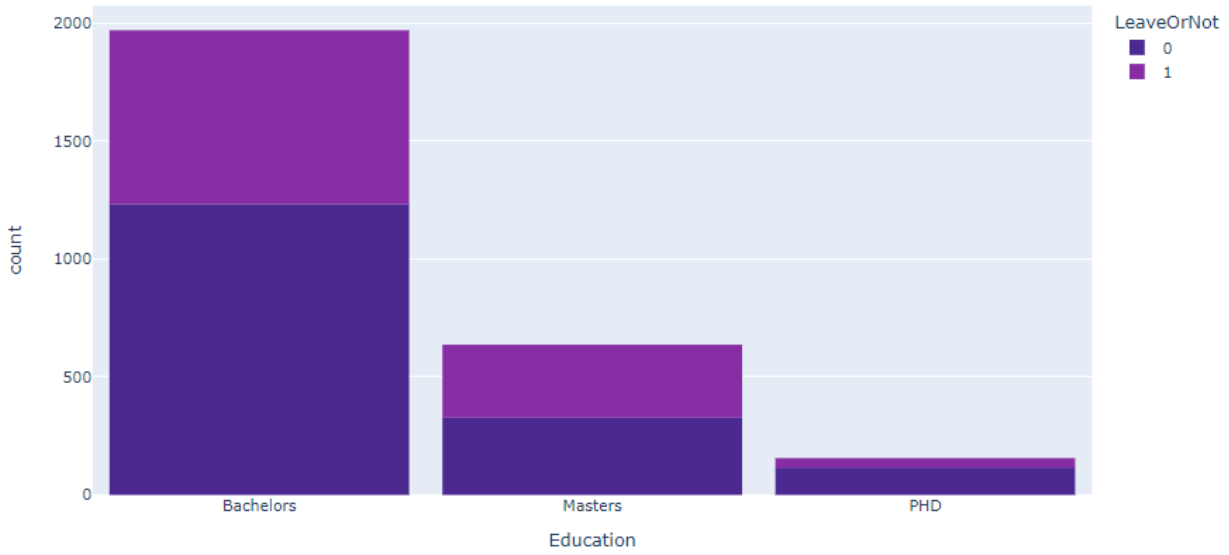
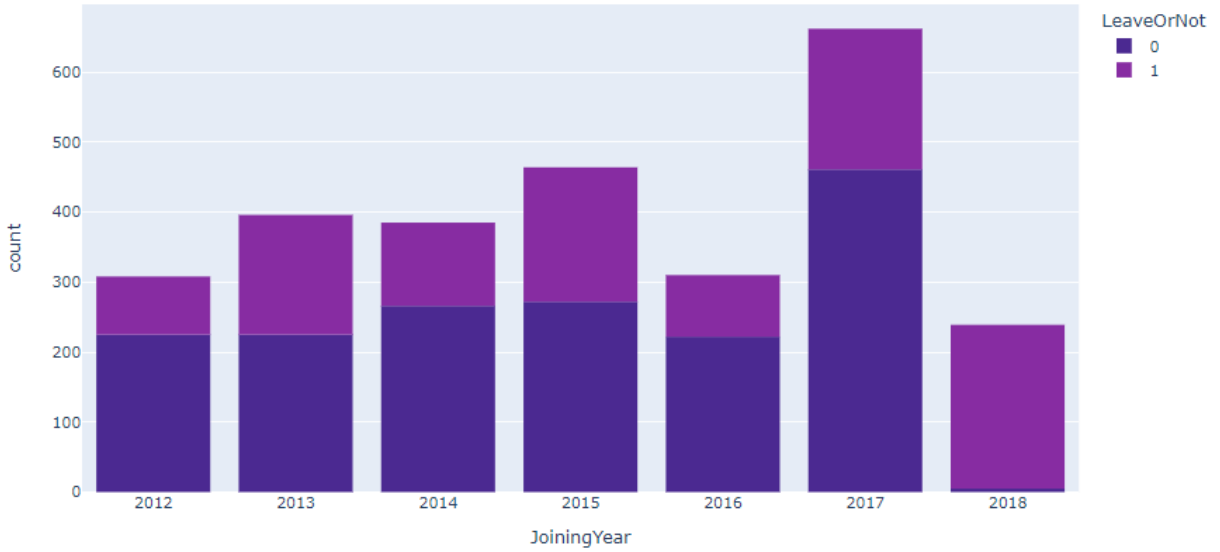


Рисунок 3 - Сравнение уровня образования

На гистограмме видно, что смена места работы зависит от уровня образования.



На гистограмме видно, что смена места работы довольно сильно зависит от года, с которого работает сотрудник. Особенно сильное влияние оказывает 2018 год.

Рисунок 4 - Сравнение по году начала работы

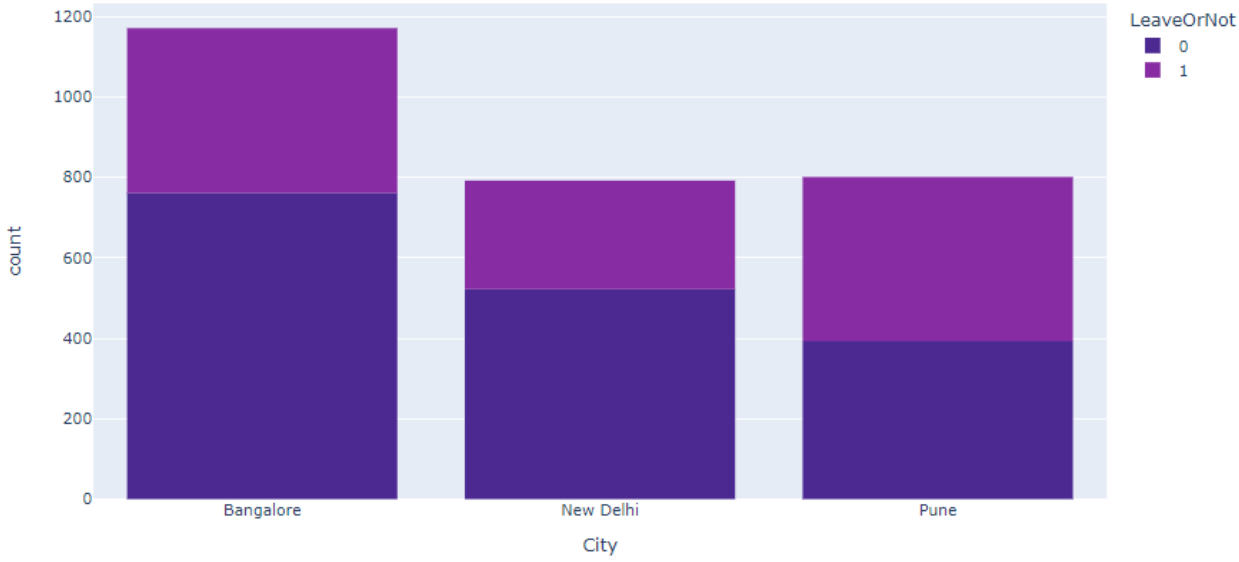


Рисунок 5 - Сравнение по городам

На гистограмме видно, что города имеют различное распределение по решениям о смене места работы.

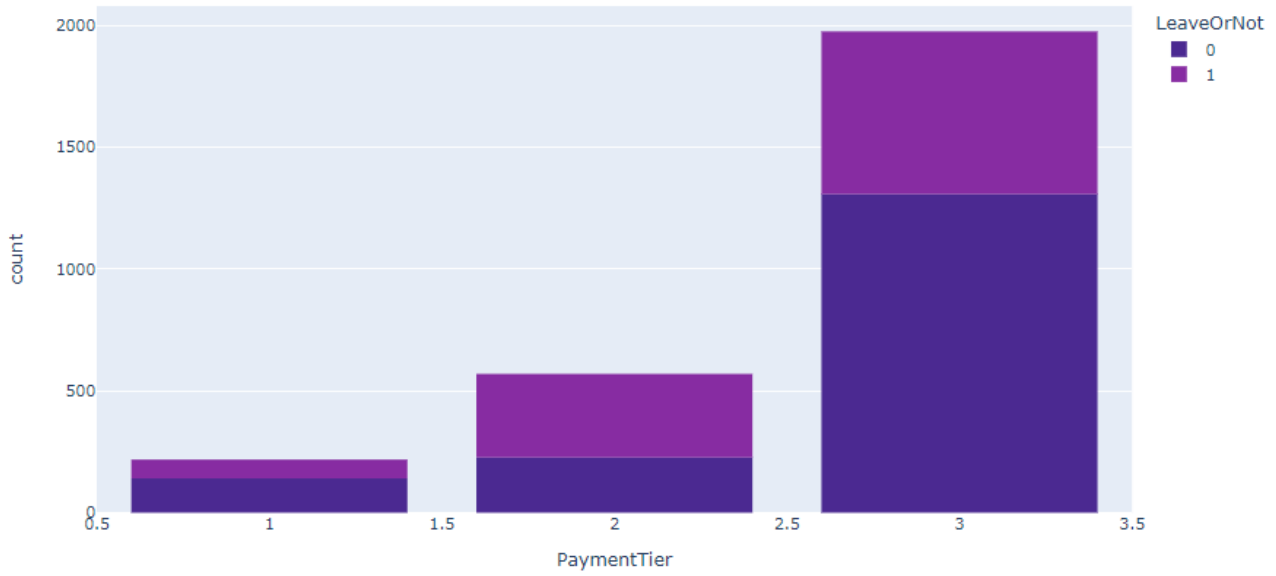


Рисунок 6 - Сравнение по уровню оплаты

На гистограмме видно, что смена решение о смене места работы имеет зависимость от уровня оплаты.

Далее приведем данные к нужному формату. Сначала масштабируем численные признаки методом Standard Scaler, который преобразует каждый признак таким образом, чтобы он имел среднее значение равное 0 и стандартное отклонение равное 1.

Затем используем OrdinalEncoder для кодирования категориальных колонок. В этом случае каждое уникальное значение признака становится новым отдельным признаком. [4]

Проводим корреляционный анализ данных. Строим тепловую карту корреляций.

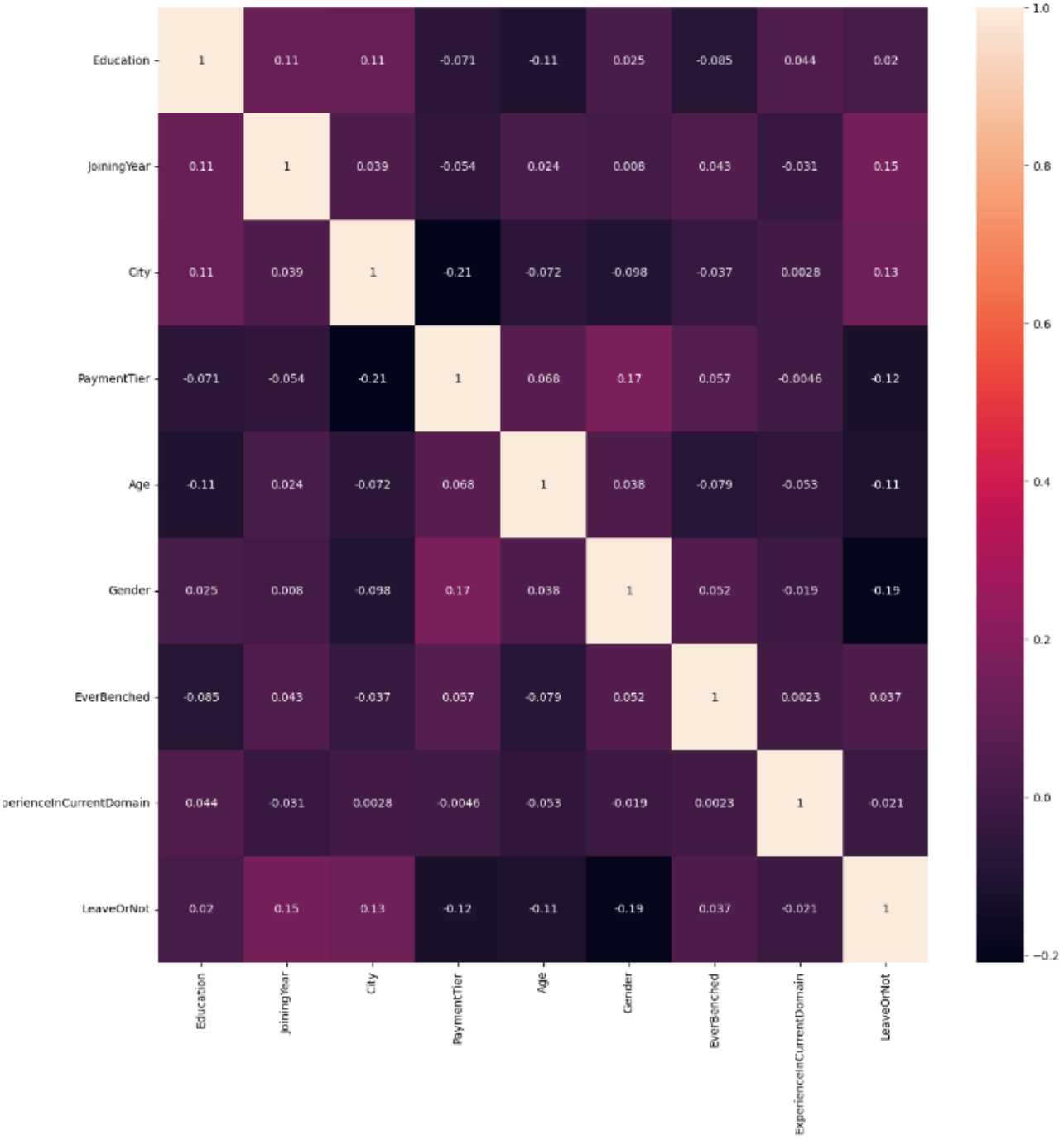


Рисунок 7 - Тепловая карта корреляций

Выводы:

- целевой признак LeaveOrNot больше всего коррелирует с возрастом (0.11), полом (0.19), городом, уровнем оплаты и годом приема;
- Образование и опыт сотрудников в текущей области на целевой признак влияют слабо;

Предварительно, по этим данным можно построить модель.

Выберем метрики для оценки качества модели:

- $Precision = \frac{TP}{TP+FP}$ - показывает, какую долю объектов, которые модель предсказала как положительные, действительно являются положительными.
- $F_1 = \frac{TP}{TP+FN}$ - показывает, какую долю положительных объектов модель способна обнаружить.
- $F_1 = 2 \times \frac{Precision \times Recall}{Precision+Recall}$ - среднее гармоническое precision и recall. Другими словами, это средневзвешенное значение точности и отзыва. [2]
- $ROC\ AUC$ - основана на вычислении следующих характеристик: $TPR = \frac{TP}{TP+FN}$ - True Positive Rate, откладывается по оси ординат. Совпадает с recall. $FPR = \frac{FP}{FP+TN}$ - False Positive Rate, откладывается по оси абсцисс. Показывает какую долю из объектов отрицательного класса алгоритм предсказал неверно. Идеальная ROC-кривая проходит через точки (0,0)-(0,1)-(1,1), то есть через верхний левый угол графика. Чем сильнее отклоняется кривая от верхнего левого угла графика, тем хуже качество классификации. [3]

  Выберем модели для решения задачи классификации:

- KNN;
- SVC;
- Дерево решений;
- Случайный лес;
- Градиентный бустинг.

  Формируем обучающую и тестовую выборку в соотношении 8:2.

  Строим базовое решения, выводим значениями метрик и ROC-кривую.

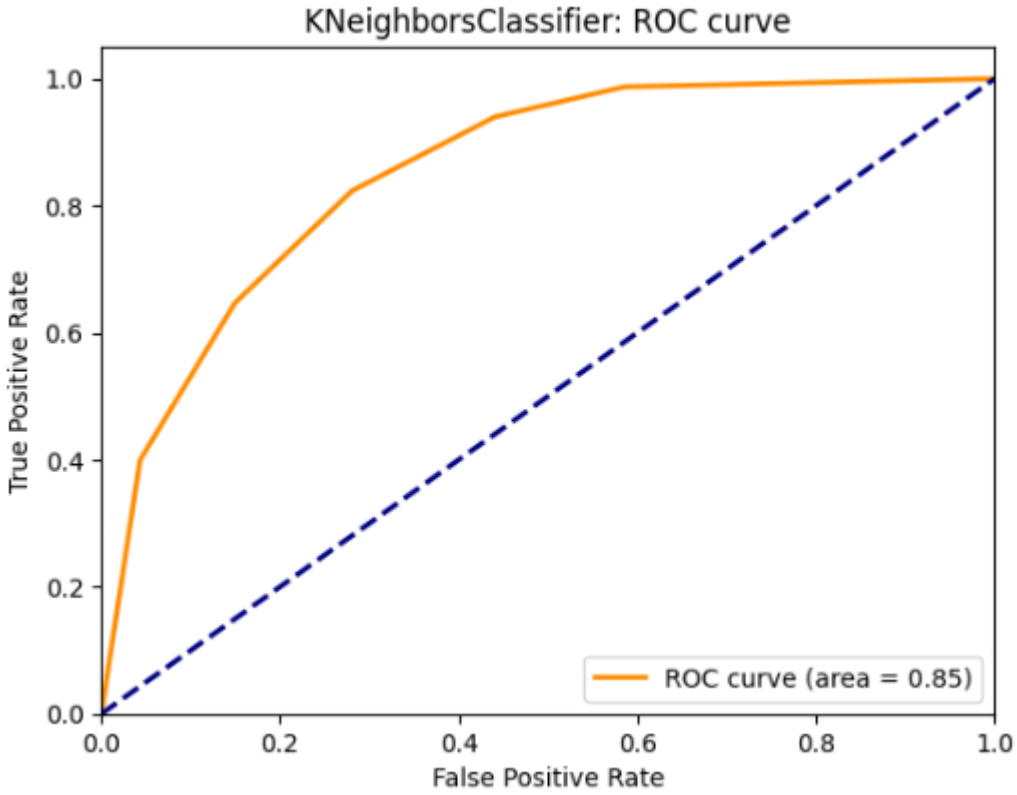


Рисунок 8 - ROC-кривая базовой модели KNN

KNeighborsClassifier:
Precision: 0.75
Recall: 0.82
F1-score: 0.79
ROC AUC score: 0.8531544653444516

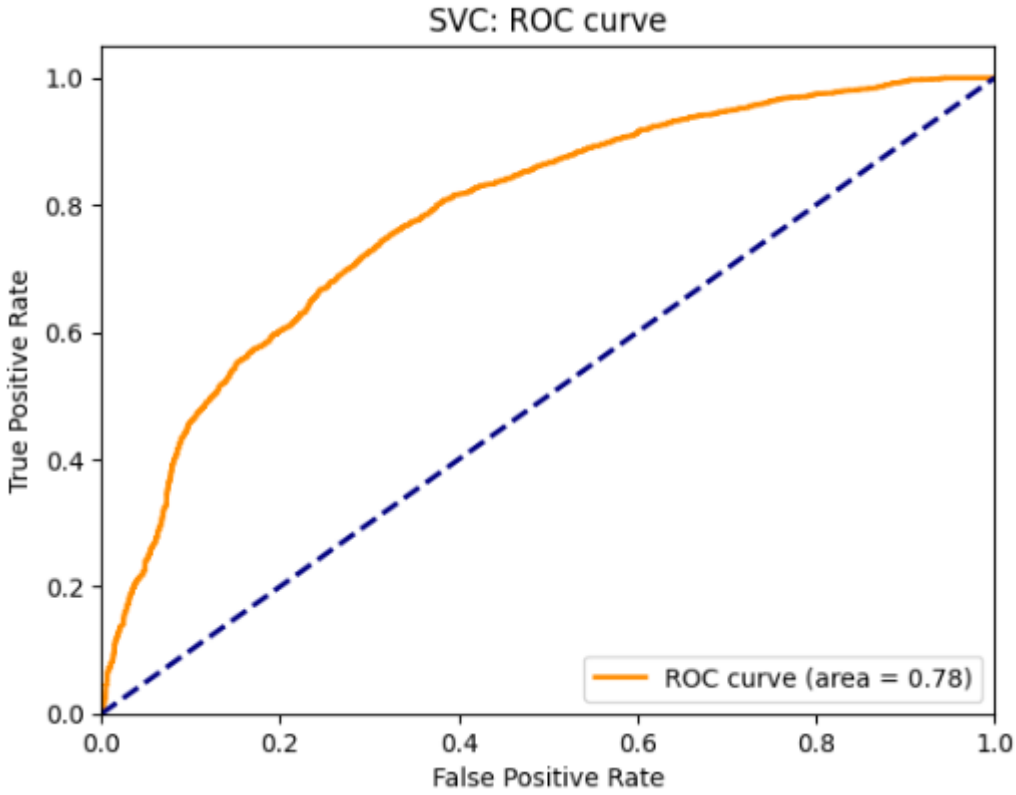


Рисунок 9- ROC-кривая базовой модели SVC

SVC:
Precision: 0.73
Recall: 0.69
F1-score: 0.71
ROC AUC score: 0.783156730530161

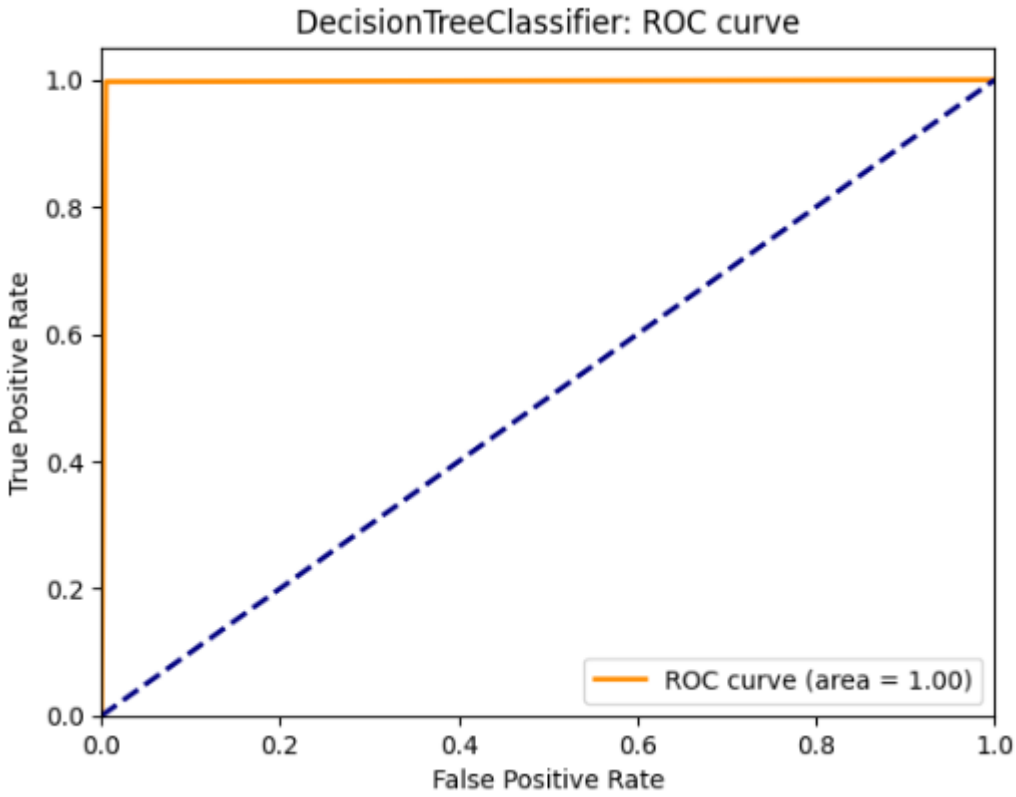


Рисунок 10 - ROC-кривая базовой модели Decision Tree

DecisionTreeClassifier:
Precision: 0.99
Recall: 1.0
F1-score: 1.0
ROC AUC score: 0.9959451681616494

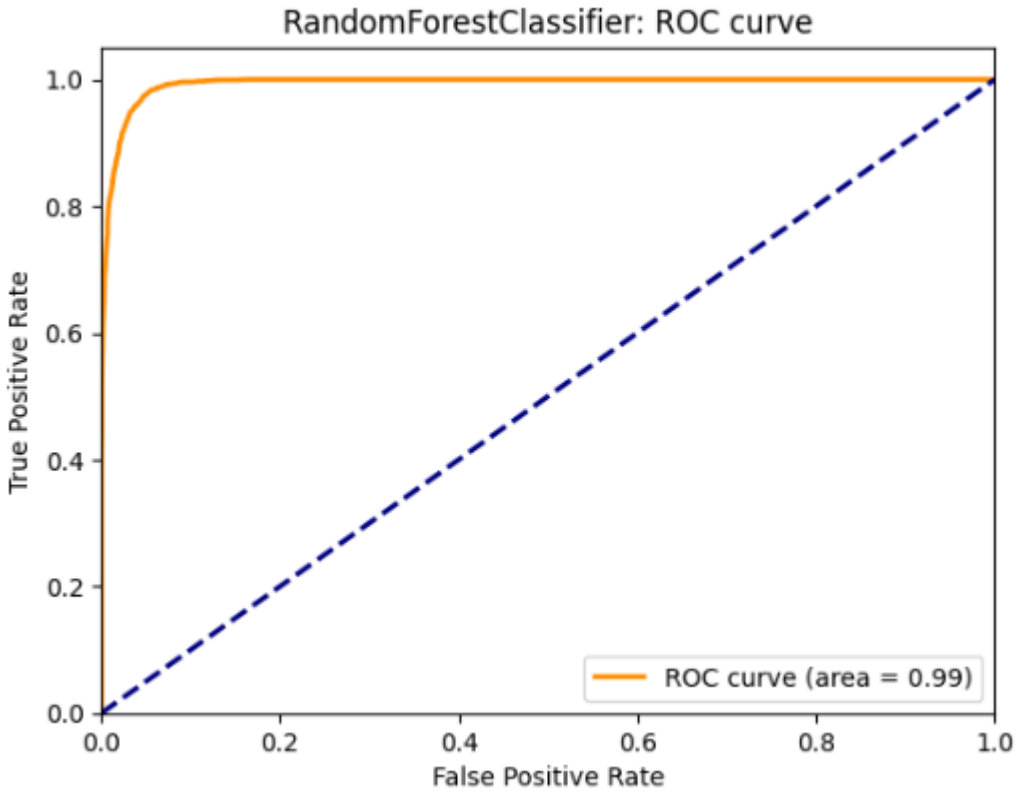


Рисунок 11 - ROC-кривая базовой модели Random Forest

RandomForestClassifier:
Precision: 0.94
Recall: 0.99
F1-score: 0.96
ROC AUC score: 0.9934043948188809

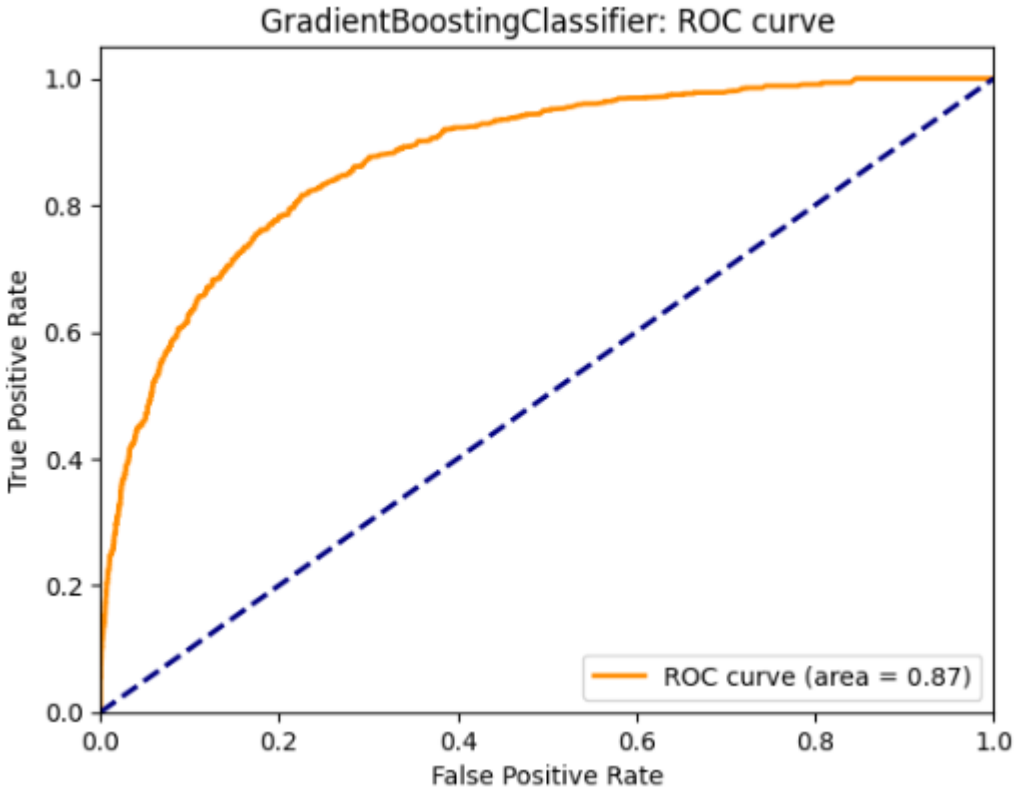


Рисунок 12 - ROC-кривая базовой модели Gradient Boosting

GradientBoostingClassifier:
Precision: 0.82
Recall: 0.76
F1-score: 0.78
ROC AUC score: 0.873846697729794

Используем GridSearch для поиска оптимальных гиперпараметров для каждой модели.

KNeighboursClassifier:

Best hyperparameters: {'algorithm': 'ball_tree', 'n_neighbors': 12, 'weights': 'uniform'}

Best score: 0.8182086216536695

SVC:

Best hyperparameters: {'C': 10, 'degree': 4, 'gamma': 'auto', 'kernel': 'rbf'}

Best score: 0.7906019223108587

DecisionTreeClassifier:

Best hyperparameters: {'criterion': 'gini', 'max_depth': 7, 'max_features': None, 'min_samples_leaf': 4, 'min_samples_split': 2}

Best score: 0.8010050764532239

RandomForestClassifier:

Best hyperparameters: {'max_depth': None, 'max_features': 'log2', 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 200}

Best score: 0.803267519892138

GradientBoostingClassifier:

Best hyperparameters: {'learning_rate': 0.05, 'max_depth': 5, 'max_features': None, 'min_samples_leaf': 2, 'min_samples_split': 5}

Best score: 0.811410273433909

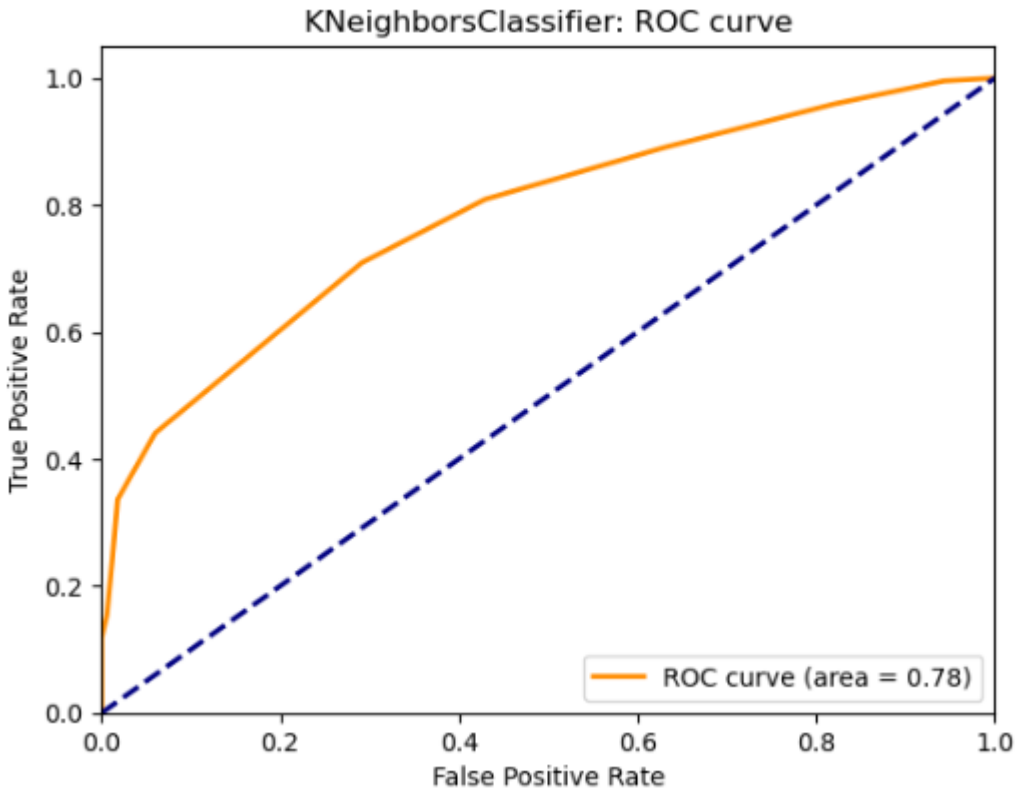


Рисунок 13 - ROC-кривая модели KNN после поиска гиперпараметров

KNeighborsClassifier:
Precision: 0.83
Recall: 0.44
F1-score: 0.58
ROC AUC score: 0.781142506142506

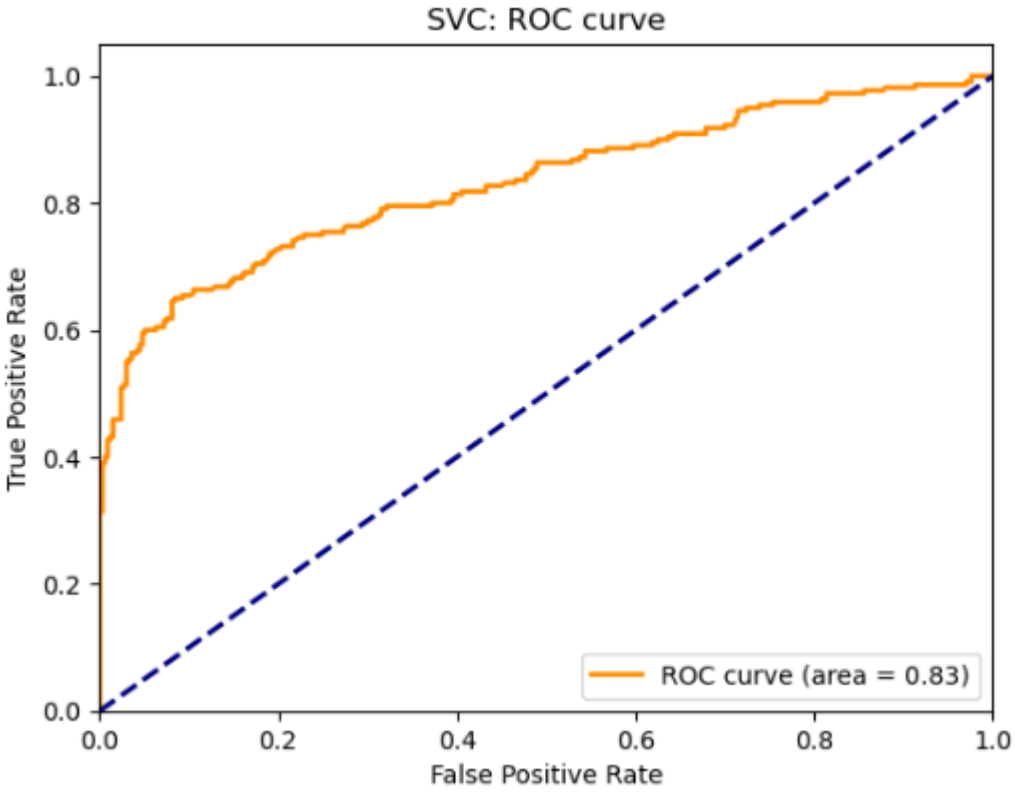


Рисунок 14 - ROC-кривая модели SVC после поиска гиперпараметров

SVC:
Precision: 0.84
Recall: 0.62
F1-score: 0.71
ROC AUC score: 0.8312653562653562

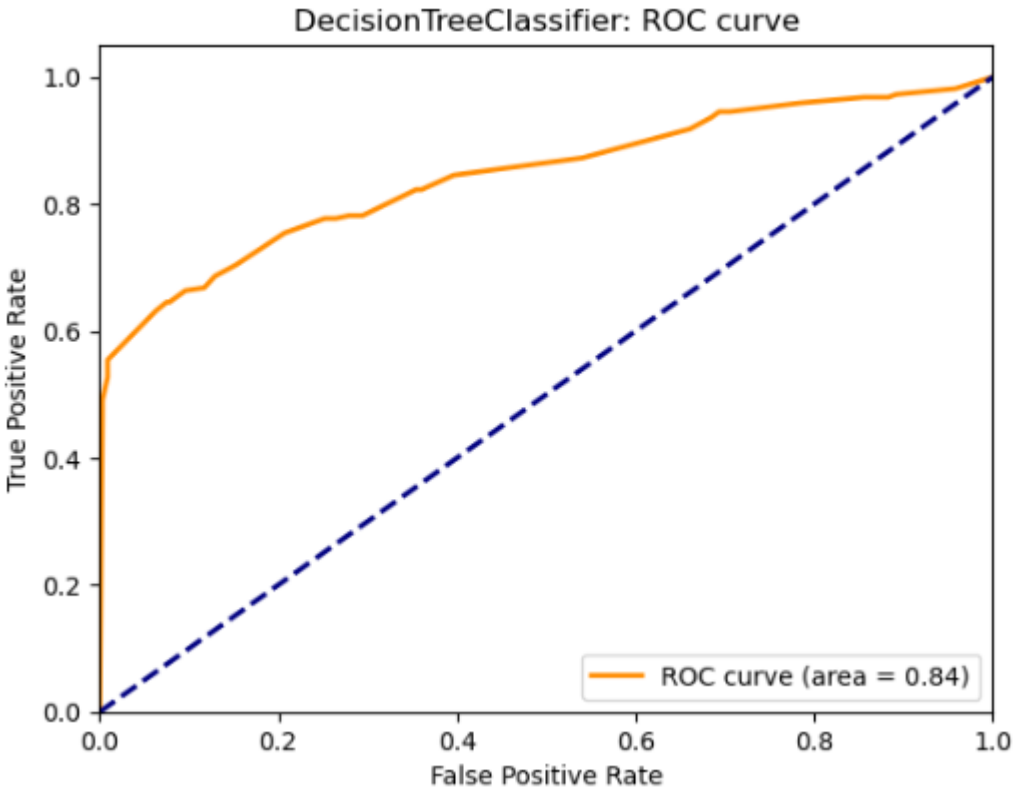


Рисунок 15 - ROC-кривая модели Decision Tree после поиска гиперпараметров

DecisionTreeClassifier:
Precision: 0.85
Recall: 0.65
F1-score: 0.73
ROC AUC score: 0.8444239694239695

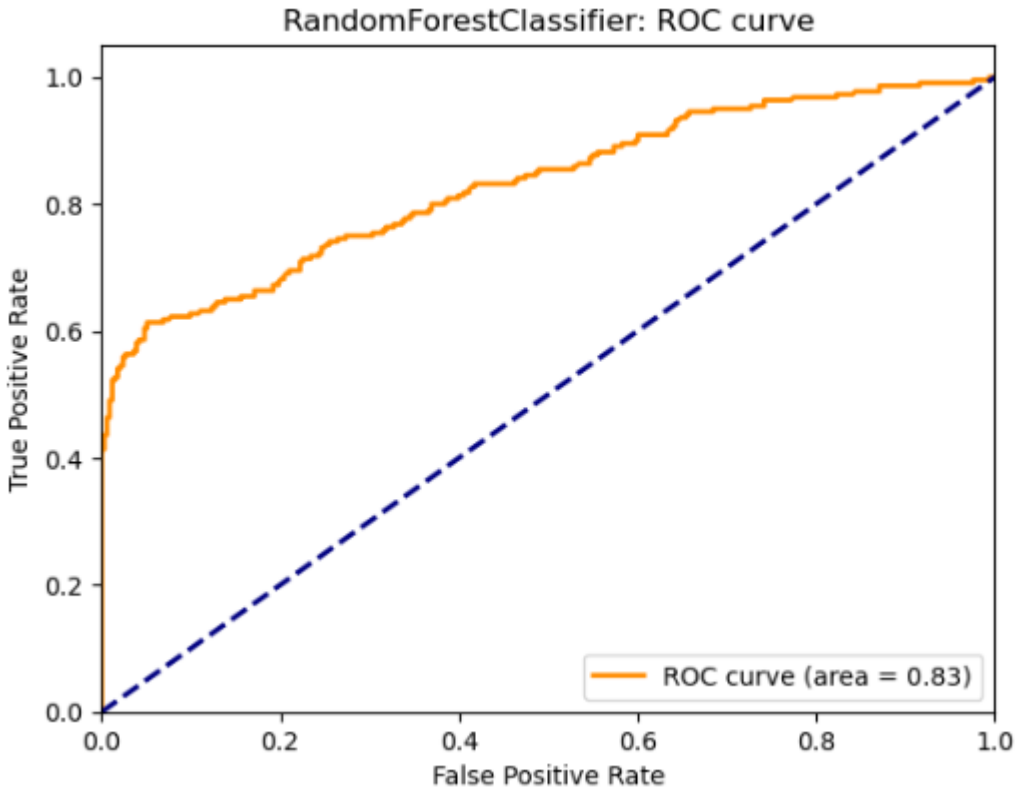


Рисунок 16 - ROC-кривая модели Random Forest после поиска гиперпараметров

RandomForestClassifier:
Precision: 0.88
Recall: 0.61
F1-score: 0.72
ROC AUC score: 0.8300778050778052

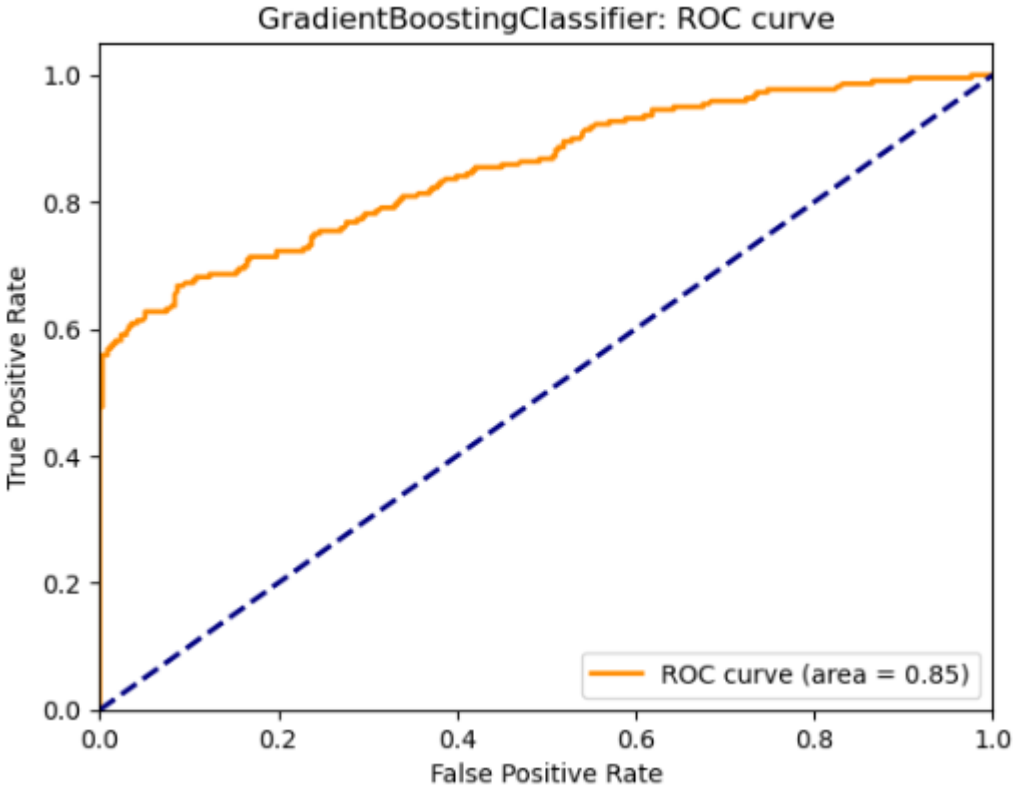


Рисунок 17 - ROC-кривая модели Gradient Boosting после поиска гиперпараметров

GradientBoostingClassifier:
Precision: 0.84
Recall: 0.63
F1-score: 0.72
ROC AUC score: 0.8529484029484029

Таблица 1 - Сравнение базовых моделей с моделями после подбора гиперпараметров по 4 метрикам

Модель	Baseline	GridSearch()
KNN	Precision: 0.75 Recall: 0.82 F1-score: 0.79 ROC AUC score: 0.8531544653444516	Precision: 0.83 Recall: 0.44 F1-score: 0.58 ROC AUC score: 0.781142506142506
SVC	Precision: 0.73 Recall: 0.69 F1-score: 0.71 ROC AUC score: 0.783156730530161	Precision: 0.84 Recall: 0.62 F1-score: 0.71 ROC AUC score: 0.8312653562653562
Decision Tree	Precision: 0.99	Precision: 0.85

	Recall: 1.0 F1-score: 1.0 ROC AUC score: 0.9959 451681616494	Recall: 0.65 F1-score: 0.73 ROC AUC score: 0.8444 239694239695
Random forest	Precision: 0.88 Recall: 0.61 F1-score: 0.72 ROC AUC score: 0.8300 778050778052	Precision: 0.88 Recall: 0.61 F1-score: 0.72 ROC AUC score: 0.8300 778050778052
Gradient Boosting	Precision: 0.82 Recall: 0.76 F1-score: 0.78 ROC AUC score: 0.8738 46697729794	Precision: 0.84 Recall: 0.63 F1-score: 0.72 ROC AUC score: 0.8529484029484029

На основании полученных метрик лучшими для решения данной задачи классификации оказались модели случайного леса и градиентного бустинга.

Заключение

Классификация решений о смене месты работы с помощью методов машинного обучения является актуальной и перспективной задачей в области кадровых движений. Анализ и обработка таких данных с помощью алгоритмов машинного обучения могут помочь в определении риска смены работы у сотрудников предприятия.

В рамках НИР была рассмотрена задача классификации сотрудников с помощью методов машинного обучения. Данные были проанализированы, визуализированы и подготовлены к обучению. Были применены различные алгоритмы, такие как метод ближайших соседей, метод опорных векторов, дерево решений, случайный лес и градиентный бустинг.

В результате исследования было показано, что большинство использованных методов могут достичь хороших результатов, но самыми точными на основании трех метрик из четырех оказались модели градиентного бустинга и случайного леса.

Список использованной литературы

1. T-test на Python для проверки и получения t-статистики // Помощник Python URL: https://pythonpip.ru/osnovy/t-test-na-python (дата обращения: 30.04.2023).

2. Machine Learning Metrics in simple terms // Medium URL: https://medium.com/analytics-vidhya/machine-learning-metrics-in-simple-terms-d58a9c85f9f6 (дата обращения: 01.05.2023).

3. Опорный пример для выполнения проекта по анализу данных. // Jupyter nbviewer URL: https://scikit-learn.org/stable/tutorial/statistical_inference/supervised_learning.html (дата обращения: 25.04.2023).

4. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр. // GitHub URL: https://github.com/ugapanyuk/courses_current/wiki/COURSE_TMO_SPRING_2023/ (дата обращения: 25.04.2023).