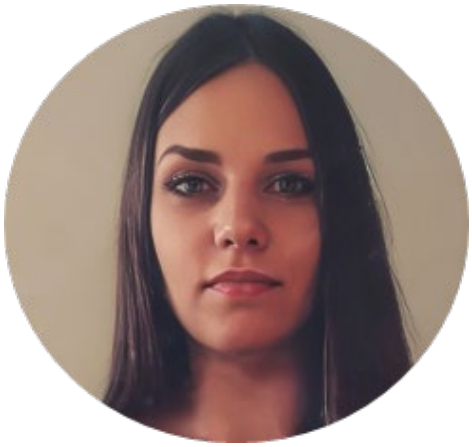

Старостина Лилия Валерьевна

г. Нерюнгри (UTC/GMT +9:00)
email: l.v.starostina2014@gmail.com
Telegram: @Lily_Val
Портфолио: https://starostinalv.github.io/
Удаленная работа - предпочтительно

Data Scientist

Стек

SQL, Python, библиотеки: Numpy, Pandas, SciPy, Statsmodels, NLTK, Re, Pymorphy2, Matplotlib, Seaborn, Sklearn, Pymystem3, Genism, Pytorch, Keras, Cv2, PIL, Pyspark, Surprise, Requests, BeautifulSoup, Time.

Ключевые компетенции

- Работа с библиотеками Python для подготовки и анализа данных;
- Написание SQL-запросов для получения данных;
- Работа с большими данными;
- Создание и обучение нейросетей;
- Работа с текстовыми данными, изображениями и временными рядами;
- Создание рекомендательных систем;
- Работа с признаками, выбор и построение моделей классического машинного обучения;
- Соблюдение дедлайнов;
- Умение обращаться со статистическими данными, применять разные методы анализа и варианты решения задач (наука, промышленные предприятия).

Профессиональный опыт

август 2022 г. – настоящее время
Учебные проекты в ООО «Нетология»
Направление «Аналитика», курс «Data Scientist» (программа курса по ссылке)

- **«Формирование SQL-запросов для анализа данных авиаперевозок»**

Цель проекта: анализ данных об авиаперевозках с помощью эффективных SQL-запросов к реляционной базе данных.
Результаты: SQL-запросы для анализа производственных и финансовых показателей в рамках задач проекта; написание SQL-запросов в нескольких вариантах и выбор наиболее оптимальных.
Технологии: PostgreSQL, DBeaver, соединения join, агрегатные функции, группировка и фильтрация, подзапросы, оконные функции, CTE, рекурсия, материализованные представления.

- **«Тематическое моделирование и классификация отзывов по тональности на основе алгоритмов классического машинного обучения»**

Цель проекта: проанализировать тексты отзывов на услуги банков, построить тематические модели; классифицировать отзывы на положительные и отрицательные.
Результаты. Выполнена обработка текстов отзывов, определены самые частотные слова (потенциальная «боль» авторов отзывов). Определены ключевые слова для положительных и отрицательных отзывов.

С помощью трех моделей сформированы наборы интерпретируемых тем, построены визуализации «Облако слов» для наглядности.

Для классификации отзывов на положительные и отрицательные проведен отбор признаков несколькими способами, на которых обучены модели логистической регрессии (с учетом дисбаланса классов). Выполнено сравнение моделей по качеству классификации с помощью метрики "F-мера", на тестовой выборке достигнуто значение 0,98.

Технологии: Python, регулярные выражения, токенизация, лемматизация, эмбеддинги, word2vec, TF-IDF, LSI, LDA, NMF, «облако слов», countvectorizer, truncatedSVD, логистическая регрессия.

- **«Перевод фраз с помощью механизма внимания»**

Цель проекта: обучить модель «seq2seq with attention» на основе конкатенации векторов, скалярного произведения и на основе MLP.

Результат: подготовлены данные (пары фраз на английском и русском языках), обучены модели «seq2seq with attention» с тремя вариантами декодеров, выполнено сравнение моделей по значению лосс-функции (минимальное достигнутое значение 1,3).

Технологии: Python, Pytorch, «seq2seq with attention», negative log likelihood loss.

- **«Детекция изображений клеток крови»**

Цель проекта: решить задачу детекции клеток крови на базе датасета «BCCD».

Результаты: обучена модель Single Shot Multibox Detection (SSD300) (за основу взята модель VGG для получения признаков), получено значение метрики Precision 0,87.

Технологии: Python, Pytorch, VGG, MultiBoxLoss, SSD.

- **«Сегментация объектов на изображении»**

Цель проекта: обучить сеть U-Net сегментировать синтетические объекты (круги) на изображении.

Результаты: подготовлены данные для обучения модели (изображения и маски), подобраны параметры модели на основе нейронной сети N-Net, проведено обучение модели. В качестве лосс-функции выбрана кросс-энтропия, на тестовой выборке значение не превышает 0,1.

Технологии: Python, Pytorch, U-Net, кросс-энтропия.

- **«Классификация изображений датасета «Cats vs Dogs»**

Цель проекта: обучить модель классифицировать изображения двух классов и достичь метрики Log Loss менее 0,3.

Результаты: за основу взяты предобученные модели VGG16 и InceptionV3, выполнено добавление слоев, подбор параметров и обучение моделей. Получены значения лосс-функции 0,23 и 0,26.

Технологии: Python, Keras, VGG16, InceptionV3.

- **«Улучшение качества обучения нейросети для классификации изображений»**

Цель проекта: применение методов Transfer Learning для улучшения качества обучения моделей классификации изображений на основе датасета «Симпсоны».

Результаты: выполнено несколько видов аугментации изображений; выполнено обучение и сравнение моделей ResNet18 с двумя типами шедулеров (ExponentialLR, MultistepLR), проведены эксперименты с сетью EfficientNet как Feature Extractor и как FineTuning. Лучшее значение Accuracy (0,94) получено при использовании сети EfficientNet как FineTuning.

Технологии: Python, Pytorch, ResNet18, EfficientNet, LR Schedulers, аугментация, сеть как Feature Extractor, сеть как Fine Tuning.

- **«Анализ временных рядов»**

Цель проекта: построение моделей временных рядов SSA временного ряда; построение моделей MA, ARIMA, GARCH, оценка ряда через HMM.

Результаты: анализ стационарных и нестационарных временных рядов. Анализ ряда среднемесячной температуры и построение предсказания с помощью модели ARIMA (с подбором оптимальных параметров по сетке). GARCH-моделирование временного ряда объемов продаж компании; его сингулярный спектральный анализ для определения составляющих ряда. Оценка ряда еженедельных значений индекса Доу-Джонса через скрытые марковские процессы.
Технологии: ARIMA, GARCH, SSA, матрица переходных вероятностей, случайные марковские процессы.

декабрь 2017 г. – май 2022 г.
Инженер-аналитик и сервисный координатор в угледобывающей промышленности (Холдинговая компания «Якутуголь», ООО «Сандвик майнинг энд констракшн СНГ») г. Нерюнгри

- Разрабатывала и вела аналитическую отчетность;
- Анализировала работу горнотранспортного и вспомогательного оборудования на основе данных первичной фактуры, систем «MODULAR MINING SYSTEMS» и «AutoGRAPH».
- Собирала данные о работе сервисной службы, вела базы «Service Work Orders» и «Сервисный модуль»; подготавливала ежемесячные отчетные документы по работе сервисной службы;
- Собирала данные и подготавливала отчеты по охране труда в системе «EHS-360»;
- Осуществляла устные и письменные переводы (английский язык).

август 2006 г. – январь 2016 г.
Работала в двух вузах (Технический институт (филиал) Северо-Восточного федерального университета в г. Нерюнгри и Новосибирском государственном техническом университете в г. Новосибирск) в должностях от ассистента до старшего преподавателя:

- Читала курсы дисциплин «Теория автоматического управления», «Метрология», «Материаловедение», «Электротехника и электроника» и др.;
- Разработала учебно-методические комплексы дисциплин;
- Руководила выпускными квалификационными работами студентов;
- Занималась научной работой (опубликовано 22 статьи, 1 монография), в частности, сбором и анализом данных для анализа энергетической безопасности Якутии.

Образование

2001 – 2006 гг. ГОУ ВПО «Якутский государственный университет им. М.К. Аммосова», Технический институт (филиал) в г. Нерюнгри, инженерный факультет. Диплом с отличием о высшем образовании, квалификация инженера по специальности «Электропривод и автоматика промышленных установок и технологических комплексов».

Дополнительное образование

- **2022 – 2025 гг.** ООО «Нетология», направление «Аналитика», курс «Data Scientist» (463 академических часа теории, 327 часов практики); программа курса по ссылке);
- **2023 г. «**Amazon Online Learning»: курсы «Math for Machine Learning», «Linear and Logistic Regression», «Data Analytics Fundamentals»;

Дополнительная информация

- Иностранные языки: английский (B2 – Upper Intermediate) китайский (A1 – Beginner)
- Водительское удостоверение категории B
- Готова к командировкам и релокации
- Хобби: спорт, комнатные цветы, иностранные языки

Обо мне

Работала в образовательной и добывающей отраслях. Изучила курс «Data science» в ООО «Нетология». Имею опыт сбора и анализа данных о работе промышленного оборудования и данных для научных исследований в области энергетики, а также опыт написания научных статей.

Работа на промышленных предприятиях дала мне понимание производственных процессов и способность принимать решения, какие данных необходимо анализировать. Благодаря участию в научных проектах получила опыт сбора, очистки и анализа статистических данных, построения и проверки гипотез, применения аналитических инструментов.

Решила продолжить профессиональное развитие в аналитике, т.к. считаю данное направление интересным, актуальным и перспективным. Наиболее интересны проекты в области обработки естественного языка, временных рядов и компьютерного зрения.