פרויקט ניהול מידע מבוזר – חלק ב׳ – Extract and Learn

רועי ברק ואלמוג בן סימון
206571135, 207375676

בהמשך לשאלה השנייה, בה היינו צריכים לאמן מודל על סט נתונים סטטי, ולחזות עליו. כעת התבקשנו לבצע זאת באמצעות נתונים מוזרמים. ראשית, מכיוון שעלינו תחילה לטעון את הנתונים וישר לבצע פרדיקציה עליהם, בנינו מודל ראשוני המבוסס על המידע הסטטי מהשאלה הקודמת. כלומר, תחילה טענו את המידע הסטטי ואימנו עליו את המודל שבחרנו בשאלה הקודמת(Random Forest עם 5 עצים בעומק של 20). לאחר מכן התחלנו בתהליך טעינת הנתונים באמצעות Streaming, כך שכל איטרציה תטען 200 אלף רשומות. ביצענו הגדרה זאת באמצעות הגדרת קונפיגורציה לאובייקט הStreaming באופן הבא ('maxOffsetsPerTrigger') :

```
streaming = spark.readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", kafka_server) \
    .option("subscribe", topic) \
    .option("startingOffsets", "earliest") \
    .option("failOnDataLoss", False) \
    .option("maxOffsetsPerTrigger", 200000) \
    .load() \
    .select(f.from_json(f.decode("value", "US-ASCII"), schema=SCHEMA).alias("value")).select("value.*")
```

ופקודה זו מגבילה את טעינת הרשומות בכל טעינה לפי הגדרת האובייקט:

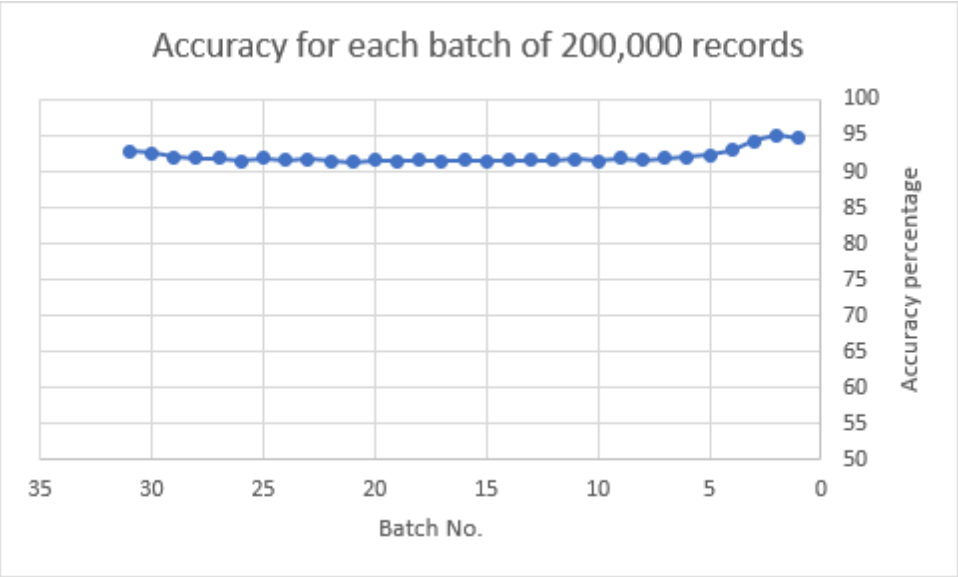
```
streaming \
    .writeStream.foreachBatch(predict_data) \
    .start() \
    .awaitTermination()
```

ביצענו את משימת החיזוי על 200 אלף רשומות בכל טעינת הנתונים באמצעות פרדיקציה עם המודל שבנינו (Random Forest עם 5 עצים בעומק של 20) בעזרת המידע הסטטי.

במהלך הרצת הקוד בשרת, נתקלנו בבעיה של שגיאת זיכרון. לכן כדי להימנע מכך הגדרנו את הקונפיגורציה של Spark להכיל 40gb בזיכרון, ובכך נמנעו משגיאת הזיכרון. להלן הקוד:

```
spark = SparkSession.builder.appName('demo_app') \
    .config("spark.kryoserializer.buffer.max", "512m") \
    .config("spark.driver.memory", "40g") \
    .getOrCreate()
```

על מנת לסיים את הStreaming נעזרנו בהגדרה שעשינו כל שבכל טעינה יטענו 200 אלף רשומות בלבד, לכן הגדרה זו נותנת לנו אינדקציה משמעותית לסוף טעינת הנתונים. לכן הוספנו תנאי כאשר נטענות פחות מ200 אלף רשומות, דרשנו להפסיק את הStreaming, ולהדפיס את ממוצע החיזוי המשוקלל, ובכך לסיים את ריצת הקוד.
לסיכום, כאשר הרצנו את הקוד שלנו בשרת הגענו לתוצאות גבוהות ולכן החלטנו שאין צורך להתאמן מחדש על הנתונים המוזרמים ולהישאר עם המודל שנבנה על ידי המידע הסטטי כדי להקל על הStreaming ולייעל את זמן הריצה. ולהלן התוצאות:



בנוסף ממוצע דיוק החיזוי על כל הרשומות הינו 92.08%.